Université de Montréal

Modèles de Markov à variables latentes : Matrice de transition non-homogène et reformulation hiérarchique

par

Gabriel Lemyre

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en mathématiques

Orientation mathématiques appliquées

4 janvier 2021

© Gabriel Lemyre, 2021

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Modèles de Markov à variables latentes : Matrice de transition non-homogène et reformulation hiérarchique

présenté par

Gabriel Lemyre

a été évalué par un jury composé des personnes suivantes :

Florian Maire

(président-rapporteur)

Maciej Augustyniak

(directeur de recherche)

Philippe Gagnon

(membre du jury)

Résumé

Ce mémoire s'intéresse aux modèles de Markov à variables latentes, une famille de modèles dans laquelle une chaîne de Markov latente régit le comportement d'un processus stochastique observable à travers duquel transparaît une version bruitée de la chaîne cachée. Pouvant être vus comme une généralisation naturelle des modèles de mélange, ces processus stochastiques bivariés ont entre autres démontré leur faculté à capter les dynamiques variables de maintes séries chronologiques et, plus spécifiquement en finance, à reproduire la plupart des *faits stylisés des rendements financiers*. Nous nous intéressons en particulier aux chaînes de Markov à temps discret et à espace d'états fini, avec l'objectif d'étudier l'apport de leurs reformulations hiérarchiques et de la relaxation de l'hypothèse d'homogénéité de la matrice de transition à la qualité de l'ajustement aux données et des prévisions, ainsi qu'à la reproduction des *faits stylisés*. Nous présentons à cet effet deux structures hiérarchiques, la première permettant une nouvelle interprétation des relations entre les états de la chaîne, et la seconde permettant de surcroît une plus grande parcimonie dans la paramétrisation de la matrice de transition. Nous nous intéressons de plus à trois extensions non-homogènes, dont deux dépendent de variables observables et une dépend d'une autre variable latente.

Nous analysons pour ces modèles la qualité de l'ajustement aux données et des prévisions sur la série des log-rendements du S&P 500 et du taux de change Canada-États-Unis (CADUSD). Nous illustrons de plus la capacité des modèles à reproduire les *faits stylisés*, et présentons une interprétation des paramètres estimés pour les modèles hiérarchiques et non-homogènes. Les résultats obtenus semblent en général confirmer l'apport potentiel de structures hiérarchiques et des modèles non-homogènes. Ces résultats semblent en particulier suggérer que l'incorporation de dynamiques non-homogènes aux modèles hiérarchiques permette de reproduire plus fidèlement les *faits stylisés*—même la lente décroissance de l'autocorrélation des rendements centrés en valeur absolue et au carré—et d'améliorer la qualité des prévisions obtenues, tout en conservant la possibilité d'interpréter les paramètres estimés.

Mots clés : Modèle de Markov à variables latentes, faits stylisés des rendements financiers, structure hiérarchique, homogénéité, variables explicatives, modèle avec dynamiques autorégressives des scores généralisées, dépendance à la durée, interprétabilité.

Abstract

This master's thesis is centered on the Hidden Markov Models, a family of models in which an unobserved Markov chain dictactes the behaviour of an observable stochastic process through which a noisy version of the latent chain is observed. These bivariate stochastic processes that can be seen as a natural generalization of mixture models have shown their ability to capture the varying dynamics of many time series and, more specifically in finance, to reproduce the *stylized facts of financial returns*. In particular, we are interested in discrete-time Markov chains with finite state spaces, with the objective of studying the contribution of their hierarchical formulations and the relaxation of the homogeneity hypothesis for the transition matrix to the quality of the fit and predictions, as well as the capacity to reproduce the *stylized facts*. We therefore present two hierarchical structures, the first allowing for new interpretations of the relationships between states of the chain, and the second allowing for a more parsimonious parameterization of the transition matrix. We also present three non-homogeneous models, two of which have transition probabilities dependent on observed explanatory variables, and the third in which the probabilities depend on another latent variable.

We first analyze the goodness of fit and the predictive power of our models on the series of log returns of the S&P 500 and the exchange rate between canadian and american currencies (CADUSD). We also illustrate their capacity to reproduce the *stylized facts*, and present interpretations of the estimated parameters for the hierarchical and non-homogeneous models. In general, our results seem to confirm the contribution of hierarchical and non-homogeneous models to these measures of performance. In particular, these results seem to suggest that the incorporation of non-homogeneous dynamics to a hierarchical structure may allow for a more faithful reproduction of the *stylized facts*—even the slow decay of the autocorrelation functions of squared and absolute returns—and better predictive power, while still allowing for the interpretation of the estimated parameters.

Keywords: Hidden Markov Model, stylized facts of financial returns, hierarchical structure, homogeneity, explanatory variables, Generalized Autoregressive Score model, duration dependence, interpretability.

Table des matières

Résumé 5

Abstract 7

Liste des spécifications de HMMs 13

Liste des algorithmes 15

Liste des tableaux 17

Table des figures 19

Liste des sigles et des abréviations 23

Abréviations 23

Notation 23

Remerciements 31

Introduction 33

Chapitre 1. Concepts de bases 35

1.1. Faits stylisés des rendements financiers 35

1.2. Distribution normale univariée 39

1.3. Modèles de mélange 42

1.3.1. Estimation des paramètres 44

Chapitre 2. Chaîne de Markov à temps discret 47

2.1. Définitions 47
2.2. Matrice de transition en m pas 49
2.3. Classification des états 49
2.4. Distribution stationnaire 50
2.5. Distributions marginales 51
Chapitre 3. Modèles de Markov à variables latentes 53
3.1. Définition 55
3.2. Distributions marginales 59
3.2.1. Distributions marginales univariées 59
3.2.2. Distributions marginales de plus grand ordre 59
3.3. Moments de $\{X_t\}_{t\in\mathbb{N}^\star}$ 61
3.4. Autocovariance de $\{X_t\}_{t\in\mathbb{N}^\star}$ et $\{C_t\}_{t\in\mathbb{N}^\star}$ 63
3.5. Log-vraisemblance 64
3.5.1. Calcul sur toutes les trajectoires possibles 64
3.5.2. Problème de filtrage 65
3.6. Inférence sur les états cachés de la chaîne 66
3.6.1. Inférence conditionnelle à $\boldsymbol{\mathcal{F}}_t$ 67
3.7. Prévision sur le processus observé 71
Chapitre 4. Reformulation hiérarchique - HHMM 75
4.1. Structure générale 77
4.1.1. Architecture de la matrice de transition 79
4.2. Structure simplifiée 82

4.2.1. Architecture de la matrice de transition 84

Chapitre 5. Extensions non-homogènes 87

5.1. Transitions dépendantes de variables observables (*Observation-driven models*) 89

5.1.1. Architecture de la matrice de transition 89

5.1.2. Formes pour les fonctions $f_{ij,t}$ 90

5.2. Transitions dépendantes de variables non-observables (*Parameter-driven models*) 95

5.2.1. *Duration-dependent Markov switching - DDMS* 95

Chapitre 6. Application 105

6.1. Qualité de l'ajustement et erreurs moyennes de prévision 107

6.2. Reproduction des faits stylisés avec les modèles entraînés sur le S&P 500 111

6.2.1. Absence d'autocorrélation dans les rendements 111

6.2.2. Présence d'autocorrélation positive dans les processus $\{|\epsilon_t|\}$ et $\{\epsilon_t^2\}$ 113

6.2.3. Regroupement des volatilités 117

6.2.4. Ailes épaisses et asymétrie négative 120

6.2.5. Effet de levier 120

6.3. Interprétation des paramètres estimés sur le S&P 500 123

6.3.1. Paramètres estimés pour le HHMM($K = 4$) 123

6.3.2. Paramètres estimés pour le HHMM-simplifié($K = 4$) 124

6.3.3. Paramètres estimés pour le HMM-LOGIT($K = 4$) $\sim$ S&P 500 125

6.3.4. Paramètres estimés pour le HMM-LOGIT($K = 4$) $\sim$ INDPRO 125

6.3.5. Paramètres estimés pour le HMM-GAS($K = 2$) 126

6.3.6. Paramètres estimés pour le HMM-DDMS($K = 4, \tau = 14$) 127

Conclusion 129

Bibliographie 131

Annexe A. Graphiques et tableaux de résultats supplémentaires 139

A.1. Résultats pour la série S&P500 140

A.2. Résultats pour la série du taux de change Canada-États-Unis (CADUSD) 151

Liste des spécifications de HMMs

1 Modèle de mélange 42

2 HMM standard 56

3 HMM généralisé 58

4 HMM Hiérarchique général 78

5 HMM Hiérarchique simplifié 83

6 *Duration-Dependent Markov Switching models* 97

Liste des algorithmes

1 Filtre d'Hamilton 68

2 Filtre d'Hamilton pour modèle *DDMS* 103

Liste des tableaux

1.1 Statistiques descriptives sur le S&P 500 38

1.2 Diverses notations pour la vraisemblance d'un échantillon $x_{1:T}$ 44

5.1 Critères de sélection de modèles basés sur la log-vraisemblance, le nombre de paramètres et la taille de l'échantillon 91

5.2 Critères de sélection de modèles basés sur l'erreur de prévision 92

5.3 Spécifications de $\Upsilon_{11}(t)$ et $\Upsilon_{22}(t)$ pour simulations 94

6.1 Résultats S&P 500 109

6.2 Résultats CADUSD 110

6.3 Statistiques descriptives sur le S&P 500 et les séries simulées 120

Table des figures

1.1 Séries des niveaux et des log-rendements du S&P 500 37

1.2 Autocorrélation des log-rendements et des log-rendements centrés en valeur absolue et au carré sur le S&P 500 39

1.3 Illustration de l'effet de levier sur les log-rendements du S&P 500 40

1.4 Densité, test des quantiles et *Box Plot* pour les log-rendements du S&P 500 [1] 41

3.1 Approche probabiliste du filtre d'Hamilton 69

5.1 Blocs sur la diagonale de $\overset{\boxtimes}{\mathbf{\Gamma}}$ pour un HMM-DDMS(K, τ) 100

5.2 Blocs hors de la diagonale de $\overset{\boxtimes}{\mathbf{\Gamma}}$ pour un HMM-DDMS(K, τ) 101

5.3 Matrice $\mathbf{\Gamma}$ sous un HMM-DDMS$(K = 4, \tau = 3)$ 102

6.1 Fonctions d'autocorrélation des log-rendements du S&P 500 112

6.2 Fonctions d'autocorrélation des log-rendements centrés en valeur absolue $\{|\epsilon_t|\}$ du S&P 500 et des séries simulées avec les modèles à l'étude 114

6.3 Fonctions d'autocorrélation des log-rendements centrés et au carré $\{\epsilon_t^2\}$ du S&P 500 et des séries simulées avec les modèles à l'étude 115

6.4 Illustration de l'effet de Taylor sur le S&P 500 et les séries simulées avec les modèles à l'étude 116

6.5 Probabilités filtrées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$ estimées sur le S&P 500 118

6.6 Probabilités lissées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_T]$ estimées sur le S&P 500 119

6.7 Illustration de l'effet de levier sur le S&P 500 et les séries simulées avec les modèles à l'étude 122

6.8 Paramètres pour le HHMM($K = 4$) estimés sur le S&P 500 123

6.9 Paramètres pour le HHMM-simplifié($K = 4$) estimés sur le S&P 500 124

A.1 Probabilités filtrées et lissées pour le modèle HMM($K = 2$) estimées sur le S&P500 140

A.2 Probabilités filtrées et lissées pour le modèle HMM($K = 3$) estimées sur le S&P500 141

A.3 Probabilités filtrées et lissées pour le modèle HMM($K = 4$) estimées sur le S&P500 142

A.4 Probabilités filtrées et lissées pour le modèle HHMM($K = 4$) estimées sur le S&P 500 143

A.5 Probabilités filtrées et lissées pour le modèle HHMM-simplifié($K = 4$) estimées sur le S&P 500 144

A.6 Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le S&P 500 avec ses valeurs décalées comme variable explicative 145

A.7 Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le S&P 500 avec la série INDPRO comme variable explicative 146

A.8 Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-GAS($K = 2$) estimées sur le S&P 500 147

A.9 Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 3$) estimées sur le S&P500 148

A.10 Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 5$) estimées sur le S&P500 149

A.11 Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 14$) estimées sur le S&P 500 150

A.12 Probabilités filtrées et lissées pour le modèle HMM($K = 2$) estimées sur le CADUSD 151

A.13 Probabilités filtrées et lissées pour le modèle HMM($K = 3$) estimées sur le CADUSD 152

A.14 Probabilités filtrées et lissées pour le modèle HMM($K = 4$) estimées sur le CADUSD 153

A.15 Probabilités filtrées et lissées pour le modèle HHMM($K = 4$) estimées sur le CADUSD 154

A.16 Probabilités filtrées et lissées pour le modèle HHMM-simplifié($K = 4$) estimées sur le CADUSD 155

A.17 Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le CADUSD avec ses valeurs décalées comme variable explicative 156

A.18 Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le CADUSD avec la série INDPRO comme variable explicative 157

A.19 Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-GAS($K = 2$) estimées sur le CADUSD 158

A.20 Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 3$) estimées sur le CADUSD 159

A.21 Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 5$) estimées sur le CADUSD 160

A.22 Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 14$) estimées sur le CADUSD 161

Liste des sigles et des abréviations

Abréviations

HMM	Modèle de Markov à variables latentes, de l'anglais *Hidden Markov Model*,
HHMM	*HMM* avec architecture hiérarchique, de l'anglais *Hierarchical Hidden Markov Model*,
GAS	Modèle avec dynamiques autorégressives des scores généralisées, de l'anglais *Generalized Autoregressive Score model*,
DDMS	Modèle de Markov à variables latentes avec dépendance à la durée, de l'anglais *Duration-Dependent Markov Switching model*.

Notation

$\mathbb{R}$	Ensemble des nombres réels,
$\mathbb{R}_+$	Ensemble des nombres réels non-négatifs, $\{x \in \mathbb{R} \vert x \geq 0\}$,
$\mathbb{R}_+^\star$	Ensemble des nombres réels positifs, $\{x \in \mathbb{R} \vert x > 0\}$,

$\mathbb{N}$	Ensemble des nombres naturels $\{0,1,2,\ldots\}$,
$\mathbb{N}^\star$	Ensemble des nombres naturels positifs $\{1,2,\ldots\}$,
$\mathbf{T}'$	Transposée de la matrice $\mathbf{T}$,
$\mathbf{x}$	Notation usuelle pour un vecteur colonne,
$\vec{x}$	Notation usuelle pour un vecteur ligne,
$\mathbf{1}_K$	Vecteur colonne de longueur K dont toutes les entrées valent 1,
$\mathbf{I}_K$	Matrice identité de dimension $K \times K$,
$\odot$	Produit d'Hadamard ou produit terme par terme,
$\otimes$	Produit de Kronecker,
$\bigoplus_{i=1}^{N} \mathbf{T}_i$	Somme directe des matrices $\mathbf{T}_i$,
$\mathrm{pa}(X)$	Les parents d'une variable X dans un graphe orienté,
$X_{t:n}$	Variables du processus $\{X_t\}_{t\in\mathbb{N}^\star}$ entre les temps t et n, soit $(X_t,X_{t+1},\ldots,X_n)$,

$x_{t:n}$	Observations associées aux variables $X_{t:n}$,
$\boldsymbol{\Theta}$	Espace des paramètres $\subset \mathbb{R}^D$,
$\boldsymbol{\theta}$	Vecteur de paramètres $(\theta_1, \ldots, \theta_D) \in \boldsymbol{\Theta}$,
$\mathcal{C}$	Espace d'états d'une chaîne de Markov,
$\boldsymbol{\Gamma}$	Notation usuelle pour la matrice de transition d'une chaîne de Markov homogène,
$\boldsymbol{\Gamma}_t$	Matrice de transition d'une chaîne de Markov non-homogène, du temps t vers le temps $t+1$,
$\boldsymbol{\Gamma}^{(m)}$	Matrice de transition en m pas d'une chaîne de Markov homogène,
$\boldsymbol{\Gamma}_t^{(m)}$	Matrice de transition en m pas d'une chaîne de Markov non-homogène, du temps t vers le temps $t+m$,
γ_{ij}	Entrée au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\boldsymbol{\Gamma}$,
$\gamma_{ij}(t)$	Entrée au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\boldsymbol{\Gamma}_t$,

$\gamma_{ij}^{(m)}$	Entrée au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\mathbf{\Gamma}^{(m)}$,
$\gamma_{ij}^{(m)}(t)$	Entrée au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\mathbf{\Gamma}_t^{(m)}$,
$\vec{\pi}$	Notation usuelle pour la distribution stationnaire d'une chaîne de Markov,
$i \longrightarrow j$	Indique que l'état j est accessible depuis l'état i,
$i \longleftrightarrow j$	Indique que les états i et j communiquent entre eux,
$p_i(x_t)$	Distribution conditionnelle $\mathbb{P}[X_t = x_t \mid C_t = i]$,
$\vec{p}(x_t)$	$(p_1(x_t), \ldots, p_K(x_t))$,
$\mathbf{P}(x_t)$	$\text{diag}(\vec{p}(x_t))$,
$p_i(x_t \mid x_{1:t-1})$	Distribution conditionnelle $\mathbb{P}[X_t = x_t \mid X_{1:t-1} = x_{1:t-1}, C_t = i]$,
$\vec{p}(x_t \mid x_{1:t-1})$	$(p_1(x_t \mid x_{1:t-1}), \ldots, p_K(x_t \mid x_{1:t-1}))$,
$\mathbf{P}(x_t \mid x_{1:t-1})$	$\text{diag}(\vec{p}(x_t \mid x_{1:t-1}))$,

$\mathfrak{u}_i(t)$	Distribution marginale du processus latent $\{C_t\}_{t\in\mathbb{N}^\star}$, $\mathbb{P}[C_t = i]$,
$\vec{\mathfrak{u}}(t)$	$(\mathfrak{u}_1(t),\ldots,\mathfrak{u}_K(t))$,
$\mathcal{L}_T(\boldsymbol{\theta})$	Vraisemblance d'une série de T observations, conditionnellement au vecteur de paramètres $\boldsymbol{\theta}$,
$\mathcal{L}_T^{-(3,4)}(\boldsymbol{\theta})$	Vraisemblance d'une série de T observations, avec les observations x_3 et x_4 manquantes, conditionnellement au vecteur de paramètres $\boldsymbol{\theta}$,
$\ell_T(\boldsymbol{\theta})$	Log-vraisemblance $\log(\mathcal{L}_T(\boldsymbol{\theta}))$,
$\ell_T^{-(3,4)}(\boldsymbol{\theta})$	Log-vraisemblance $\log\Big(\mathcal{L}_T^{-(3,4)}(\boldsymbol{\theta})\Big)$,
$\tilde{\mu}_p$	Moment centré d'ordre p, soit $\mathbb{E}\Big[\Big(\frac{X-\mu}{\sigma}\Big)^p\Big]$,
μ	Espérance inconditionnelle $\mathbb{E}[X_t]$,
μ_i	Espérance conditionnelle $\mathbb{E}[\,X_t \mid C_t = i\,]$,
$\mu_{g\|i}$	Espérance conditionnelle $\mathbb{E}[\,g(X_t) \mid C_t = i\,]$,
$\boldsymbol{\mu}_g$	$\Big(\mu_{g\|1},\ldots,\mu_{g\|K}\Big)'$,

$\mathbf{M}_g$	$\text{diag}\Big(\boldsymbol{\mu}_g\Big)$,
σ^2	Variance inconditionnelle $\mathbb{V}ar[X_t]$,
σ_i^2	Variance conditionnelle $\mathbb{V}ar[X_t \mid C_t = i]$,
$\sigma_{g\|i}^2$	Variance conditionnelle $\mathbb{V}ar[g(X_t) \mid C_t = i]$,
$\boldsymbol{\sigma}_g^2$	$\Big(\sigma_{g\|1}^2, \ldots, \sigma_{g\|K}^2\Big)'$,
$\boldsymbol{\Sigma}_g^2$	$\text{diag}\Big(\boldsymbol{\sigma}_g^2\Big)$,
${}_{g_1,g_2}\rho_{X_t}^{(h)}$	Autocorrélation $\mathbb{C}\text{orr}[g_1(X_t), g_2(X_{t+h})]$,
$\vec{a}_t$	$(\mathbb{P}[X_t,C_t = 1 \mid \boldsymbol{\mathcal{F}}_{t-1}], \ldots, \mathbb{P}[X_t,C_t = K \mid \boldsymbol{\mathcal{F}}_{t-1}])$,
a_t	$\mathbb{P}[X_t \mid \boldsymbol{\mathcal{F}}_{t-1}]$,
$\vec{\omega}_t$	$(\mathbb{P}[C_t = 1 \mid \boldsymbol{\mathcal{F}}_t], \ldots, \mathbb{P}[C_t = K \mid \boldsymbol{\mathcal{F}}_t])$,
$\vec{\omega}_{t+h \mid t}$	$(\mathbb{P}[C_{t+h} = 1 \mid \boldsymbol{\mathcal{F}}_t], \ldots, \mathbb{P}[C_{t+h} = K \mid \boldsymbol{\mathcal{F}}_t])$,
$\vec{\omega}_{t+h \mid t}(i)$	$\mathbb{P}[C_{t+h} = i \mid \boldsymbol{\mathcal{F}}_t]$,

$\overset{\triangle}{\mathcal{C}}$	Espace d'états de la couche supérieure d'un *HMM* hiérarchique,
$\overset{\circ}{\mathcal{C}}$	Espace d'états de la couche subalterne d'un *HMM* hiérarchique,
$\overset{\triangle}{\mathbf{\Gamma}}$	Matrice de transition de la couche supérieure d'un *HMM* hiérarchique,
$\overset{\circ}{\mathbf{\Gamma}}_i$	Matrice de transition du i^{e} régime de la couche subalterne d'un *HMM* hiérarchique,
$\overset{\triangle}{\gamma}_{ij}$	Entrée au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\overset{\triangle}{\mathbf{\Gamma}}$,
$\overset{\circ}{\gamma}{}^{i}_{\eta\ell}$	Entrée au croisement de la η^{e} ligne et de la ℓ^{e} colonne de la matrice $\overset{\circ}{\mathbf{\Gamma}}_i$,
ϖ^i_ℓ	Probabilité de choisir le ℓ^{e} régime subalterne du i^{e} régime supérieure d'un *HMM* hiérarchique,
s^i_ℓ	Probabilité de quitter le ℓ^{e} régime subalterne du i^{e} régime supérieure d'un *HMM* hiérarchique,
$\mathcal{S}$	Espace d'états de la chaîne *instrumentale* d'un HMM-DDMS(K,τ),
D_t	Durée de temps dans un état d'un modèle HMM-DDMS(K,τ),

$\overset{\blacktriangledown}{\mathbf{\Gamma}}$	Matrice de transition de la chaîne *instrumentale* d'un HMM-DDMS(K,τ),
$\overset{\blacktriangledown}{\gamma}_{ij}$	Entrée au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\overset{\blacktriangledown}{\mathbf{\Gamma}}$,
$\overset{\blacktriangledown}{\boldsymbol{\alpha}}(1)$	Distribution marginale du processus latent de la chaîne *instrumentale* d'un HMM-DDMS(K,τ),
$\overset{\blacktriangledown}{p}(x_t \mid x_{1:t-1})$	Vecteur des distributions conditionnelles de X_t dans la chaîne *instrumentale* d'un HMM-DDMS(K,τ).

Remerciements

J'aimerais d'abord remercier mon directeur de recherche Maciej Augustyniak. Je tiens particulièrement à le remercier pour sa patience et son empathie tout au long de ce processus qui n'aura certe pas été évident. Il aura été un allié inestimable et un conseiller sans pareil lors de la rédaction de ce mémoire et de mon parcours à l'université en général jusqu'à maintenant.

Je tiens aussi à remercier ma famille, sans qui rien de ceci ne serait possible. Je remercie mes parents qui m'ont supporté depuis le début, qui m'ont aidé à braver toutes les tempêtes et qui m'ont inculqué l'importance du travail bien fait. Je remercie de plus mon grand-père, qui m'a beaucoup appris sur le milieu de la finance et qui était toujours disponible pour répondre à mes questions. Je remercie aussi mes deux soeurs pour leur support durant ce défi, et bien d'autres avant lui.

Je remercie aussi tou.te.s mes ami.e.s et collègues à l'université sans qui cette expérience n'aurait pas été la même. En particulier, j'aimerais remercier Estelle Filiatrault-Bégin pour son soutien autant morale qu'alimentaire durant cette épreuve, ainsi que Thomas Davignon et Alexandre Paul pour leur aide, leur humour et leur compagnie durant les longues séances de rédaction.

J'aimerais de plus remercier toutes mes colocataires qui m'ont épaulé durant cette épreuve. Leur soutien et leurs conseils judicieux quant au maintien de ma santé mentale ont certainement fait la différence.

Je remercie aussi tous les membres de l'AECSMS et spécialement ceux et celles avec qui j'ai eu la chance de siéger au conseil exécutif. Cette expérience aura été très importante dans mon parcours et la camaraderie que j'y ai trouvé aura certainement été un des facteurs prépondérants pour ma réussite académique.

Je tiens de plus à remercier l'équipe *R Core*[2] puisque toutes les analyses présentées dans ce mémoire furent effectuées grâce au logiciel `R` et aux librairies s'y trouvant. De plus, la

2. R Core Team [2013].

totalité de ce mémoire fût rédigé grâce au langage LaTeX créé par Leslie Lamport [3] et qui existe grâce à l'incroyable travail de Donald E. Knuth [4], créateur du langage TeX. Je les remercie chaleureusement, ainsi que leurs collaborateurs.

Finalement, je remercie le Fonds de recherche québécois Nature et technologies (FRQNT), le département de mathématiques et statistique (DMS) de l'Université de Montréal ainsi que mon directeur Maciej Augustyniak pour leur soutien financier.

3. Lamport [1986].
4. Knuth [1984].

Introduction

Un modèle n'étant autre chose en fait qu'une série d'hypothèses, il convient, lorsque l'on étudie un modèle, de débuter par une analyse de ces hypothèses, ainsi que des conséquences de celles-ci.

À cet égard, le présent mémoire s'intéresse aux modèles de Markov à variables latentes, une famille de modèles dans laquelle un phénomène est représenté par une chaîne de Markov latente dont seule une version buitée est observée. Les modèles de cette famille ont connu une popularité grandissante en finance depuis les travaux de Hamilton [1989]. Ceci n'est pas surprenant puisque plusieurs caractéristiques de ces modèles en font d'excellents candidats à la modélisation de séries financières. En bâtissant sur les travaux de Granger et Ding [1995], l'article de Rydén, Teräsvirta et Åsbrink [1998] montrait par exemple que les modèles de Markov à variables latentes se démarquent en ce qui a trait à la reproduction des *faits stylisés des rendements financiers*. Ces *faits stylisés*, caractéristiques observables sur une grande quantité de séries financières, doivent être reproduits aussi fidèlement que possible par les modèles statistiques utilisés sur ces dernières. Dans cet article, ces derniers auteurs illustrent que les modèles de Markov à variables latentes homogènes ayant des distributions normales comme noyau ont la capacité de reproduire tous ces faits stylisés à l'exception du second, soit la lente décroissance de l'autocorrélation des rendements, en valeur absolue et au carré.

À l'instar de ces derniers, la majorité de la littérature au sujet des modèles de Markov à variables latentes suppose que la matrice de transition de la chaîne de Markov est homogène. Puisque l'homogénéité d'une chaîne de Markov permet de multiples simplifications, il est évident que sans cette propriété, plusieurs quantités d'intérêt sont plus complexes—voir impossibles—à obtenir. En plus d'alourdir la notation, la non-homogénéité signifie de plus que la matrice de transition change potentiellement à chaque pas, ce qui peut appesantir considérablement la routine d'optimisation.

Cette hypothèse très populaire étant cependant très forte, nous nous intéresserons dans ce mémoire à l'apport potentiel d'une relaxation de celle-ci à la qualité de l'ajustement du modèle aux données et à sa faculté à reproduire les *faits stylisés des rendements financiers*.

L'interprétabilité des paramètres étant souvent importante dans les applications en finance, celle-ci sera aussi une préoccupation tout au long de notre étude. Les modèles non-homogènes étant généralement plus lourdement paramétrés et la parcimonie étant une caractéristique convoitée lors de la sélection de modèles, nous nous intéresserons par le fait même à une reformulation hiérarchique du modèle permettant une plus grande parcimonie, mais aussi une nouvelle interprétation des relations entre les états.

Les résultats obtenus lors de nos analyses suggèrent d'ailleurs que la combinaison d'une reformulation hiérarchique et d'une structure non-homogène puisse permettre au modèle résultant de reproduire plus fidèlement les faits stylisés—même le second mentionné ci-haut—, et de produire des prévisions plus justes sur la série à l'étude. À ce propos, cette série étant très populaire dans les articles ayant trait aux modèles de Markov à variables latentes —et dans la littérature financière en général—nous utiliserons les log-rendements de l'indice S&P 500 aux fins d'illustration des concepts et de test de nos modèles.

La suite de ce mémoire sera structurée de la manière suivante ; la revue de littérature propre à chacun des chapitres sera présentée en introduction de ceux-ci. Au chapitre 1 (p.35), nous présenterons certains concepts de base liés à l'étude de séries financières—log-rendements et faits stylisés—, illustrerons ces concepts sur le S&P 500 et présenterons les modèles de base n'incluant pas de dépendance temporelle—distributions normales univariées et modèles de mélange. Au chapitre 2 (p.47), nous introduirons les chaînes de Markov ainsi que les principales caractéristiques de ces dernières puis présenterons au chapitre 3 (p.53) les modèles de Markov à variables latentes, combinant pour ainsi dire les modèles de mélange aux chaînes de Markov. Nous nous intéresserons ensuite, au chapitre 4 (p.75), à deux reformulations hiérarchiques de ces modèles, la première permettant une nouvelle interprétation des relations entre les états de la chaîne de Markov, et la seconde permettant de surcroît une plus grande parcimonie dans la construction de la matrice de transition. Au chapitre 5 (p.87), nous présenterons deux catégories d'extensions non-homogènes des modèles de Markov à variables latentes. D'abord, nous présenterons deux extensions pour lesquelles les probabilités de transition dépendent de variables observables—soit le HMM-LOGIT(K) et le HMM-GAS(K) —et ensuite, une extension où elles dépendent d'une variable latente—le HMM-DDMS(K,τ)—permettant d'inclure le nombre de pas consécutifs dans un même état aux probabilités de transition. Nous conclurons ce mémoire au chapitre 6 (p.105) par une étude des performances de ces modèles sur la séries des log-rendements du S&P 500. En particulier, nous étudierons la qualité de l'ajustement—log-vraisemblance, *AIC*, *AICc* et *BIC*—, l'erreur moyenne de prévision—*MAFE*, *MSFE*, *MASFE* et *MSSFE*—, la capacité de ces modèles à reproduire les faits stylisés et finalement l'interprétation des paramètres estimés.

Chapitre 1

Concepts de bases

1.1. Faits stylisés des rendements financiers

Puisque nous nous intéresserons à l'évolution du niveau d'un indice, ou du prix d'un titre, il est d'abord convenable de définir la forme que prendra la relation entre les prix successifs et les log-rendements associés. En notant P_t le prix au temps t, nous modéliserons la relation entre ce prix et P_{t-1} par le biais de la variable aléatoire $r_t \in \mathbb{R}$. Cette relation prend la forme

$$P_t = P_{t-1}e^{r_t} \Leftrightarrow r_t = \log\left(\frac{P_t}{P_{t-1}}\right) = \log(P_t) - \log(P_{t-1}). \tag{1.1.1}$$

Cette définition pour r_t correspond à un rendement composé continûment entre $t-1$ et t.

Lorsque l'on tente de modéliser ces rendements financiers, nous sommes confrontés à plusieurs dynamiques complexes, difficiles à reproduire avec des modèles économétriques standards. Puisqu'elles sont communes à la majorité des séries financières, quelques unes de ces dynamiques sont rassemblées sous l'appelation des *faits stylisés des rendements financiers*.

Il y a bien longtemps que les économétriciens s'intéressent à ces caractéristiques. Lorsque vient le temps de justifier l'architecture d'un modèle, par exemple, ces hypothèses doivent au mieux être reproduites par celui-ci. Certaines de ces caractéristiques étaient même déjà étudiées dans les travaux révolutionnaires de Louis Bachelier au début du 20^{e} siècle, de Osborne [1959] et de Fama [1965]. Ce dernier affirmait alors, à contre-courant des « traders » de l'époque, qu'il n'y avait aucune preuve de l'existence de dépendance dans le processus des rendements, mais que de grandes variations dans les rendements avaient tendance à être suivies de grandes variations à leur tour.

Ce qui était presque scandaleux à l'époque ne l'est plus maintenant. En effet, ces *faits stylisés* étant très importants, ils ont été étudiés depuis par de multiples auteurs comme Boothe et Glassman [1987], Diebold [1988], Engel et Hamilton [1990], Rydén, Teräsvirta

et Åsbrink [1998], Maheu et McCurdy [2000b], Ang et Bekaert [2002b], Augustyniak [2018] et bien d'autres. La manière d'exposer ces *faits stylisés* diffère quelque peu en fonction de l'époque ou de l'auteur, mais les idées restent sensiblement les mêmes. Nous présenterons un mélange des formulations de Rydén, Teräsvirta et Åsbrink [1998] et de Augustyniak [2018].

Afin de bien illustrer ces propriétés des rendements financiers, nous adopterons la forme suivante pour les log-rendements :

$$r_t = \mu_t + \epsilon_t = \mu_t + \eta_t \sigma_t, \tag{1.1.2}$$

avec

$$\mu_t = \mathbb{E}[r_t \mid r_{1:t-1}], \tag{1.1.3}$$

$$\sigma_t^2 = \mathbb{V}ar[r_t \mid r_{1:t-1}], \tag{1.1.4}$$

$$\eta_t \sim \mathcal{N}(0,1) \tag{1.1.5}$$

et le vecteur des log-rendements passés noté $r_{1:t-1} \stackrel{\text{def}}{=} (r_1, r_2, \ldots, r_{t-1})$. Le présent mémoire étant principalement dédié à l'analyse de séries chronologiques, la notation $X_{1:n} \stackrel{\text{def}}{=} (X_1, \ldots, X_n)$ correspondante à un vecteur de n variables aléatoires indicées et ordonnées sera utilisée à maintes reprises dans les sections qui suivent.

Les faits stylisés des rendements financiers se résument en cinq points :

(1) Absence d'autocorrélation dans les rendements (i.e. $\mathbb{C}\text{orr}[r_t, r_{t+h}] \approx 0, \forall h \geq 1$). [1]

(2) Présence d'autocorrélation positive dans les processus $\{\epsilon_t^2\}$ et $\{|\epsilon_t|\}$, ce qui implique de la dépendance entre les rendements centrés, mais surtout une hétéroscédasticité conditionnelle (i.e. $\sigma_t^2 \neq \sigma^2$). Cette dépendance est observable en calculant $\mathbb{C}\text{orr}\left[\epsilon_t^2, \epsilon_{t+h}^2\right]$ et $\mathbb{C}\text{orr}[|\epsilon_t|, |\epsilon_{t+h}|]$. De plus, la décroissance de ces autocorrélations est bien plus lente que celle associée à un processus AR(1) ou ARMA(1,q). [2]

(3) Regroupement (*clustering*) des volatilités : Les rendements de faibles et de fortes amplitudes ont tendance à se regrouper, ce qui est dû au point (2).

(4) Ailes épaisses et asymétrie négative : Les rendements auront un coefficient d'aplatissement [3] (*kurtosis*) supérieur à celui d'une densité normale (i.e. $\tilde{\mu}_4 > 3$), ce qui

1. La corrélation entre deux variables aléatoires quelconques X et Y est définie comme
$$\mathbb{C}\text{orr}[X,Y] = \frac{\mathbb{C}ov[X,Y]}{\sqrt{\mathbb{V}ar[X]}\sqrt{\mathbb{V}ar[Y]}} = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{V}ar[X]}\sqrt{\mathbb{V}ar[Y]}}$$

2. Rydén, Teräsvirta et Åsbrink [1998].

3. Coefficient d'aplatissement : $\tilde{\mu}_4 = \mathbb{E}\left[\left(\frac{r-\mu}{\sigma}\right)^4\right]$

implique des ailes de distribution épaisses. De plus, ils auront un coefficient d'asymétrie [4] (*skewness*) négatif ($\tilde{\mu}_3 < 0$), indiquant une prépondérance de rendements positifs, mais des rendements négatifs plus extrêmes.

(5) Effet de levier : Un choc négatif sur les rendements aura un effet plus fort sur la volatilité des rendements subséquents qu'un choc positif de la même amplitude. Mathématiquement, cet effet s'écrit $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{-\epsilon_t, 0\}] > \mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{\epsilon_t, 0\}]$.

Remarque 1.1.1. Selon Rydén, Teräsvirta et Åsbrink [1998], le fait sylisé le plus difficile à reproduire—et le plus important pour la modélisation—est la lente décroissance de l'autocorrélation dans les processus $\{\epsilon_t^2\}$ et $\{|\epsilon_t|\}$, soit le fait stylisé (2). L'hypothèse soulevée par les auteurs pour expliquer cette difficulté est que les modèles de Markov à variables latentes standards sont, par construction, incapables de tenir compte de tendances à long terme dans les rendements. Nous présenterons une solution possible à ce problème au chapitre 5 (p.87), soit l'utilisation de chaînes de Markov non-homogènes.

Tel que mentionné en introduction, nous utiliserons les log-rendements sur l'indice S&P 500 afin d'illustrer le processus nous menant à introduire les modèles de Markov à variables latentes. Illustrons donc d'abord les faits stylisés par le biais de cette série afin de bien cerner de quoi il est question. La figure 1.1 présente la série sans transformation ainsi que ses log-rendements. Au premier coup d'oeil sur le deuxième graphique de la figure 1.1, nous pouvons

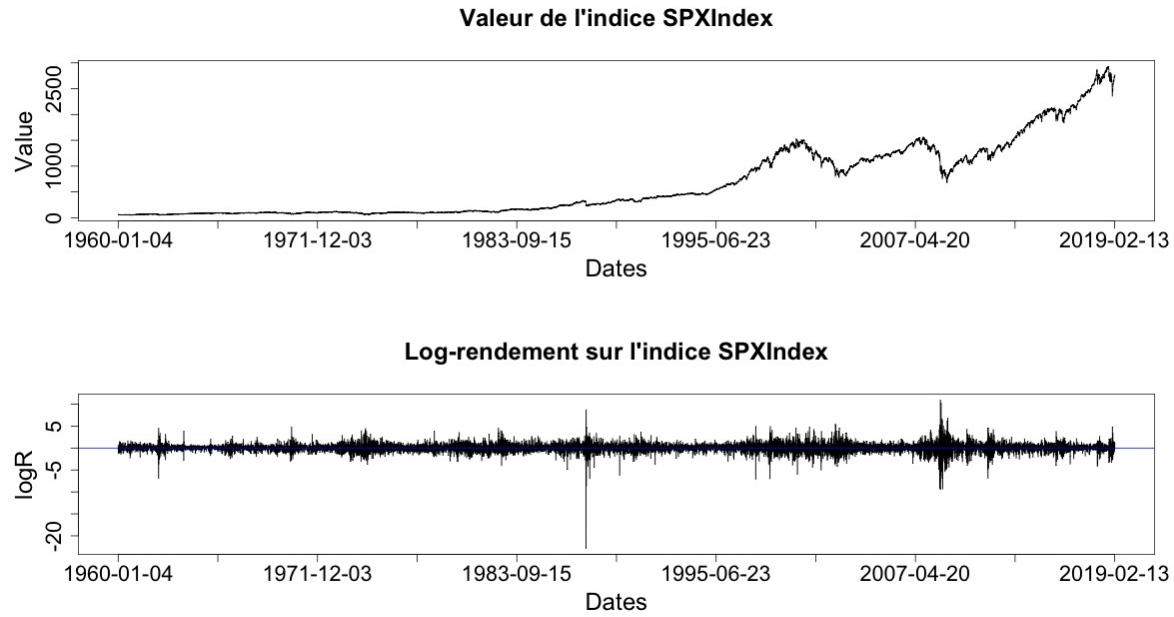


Figure 1.1 – Séries des niveaux et des log-rendements du S&P 500

4. Coefficient d'asymétrie : $\tilde{\mu}_3 = \mathbb{E}\left[\left(\frac{r-\mu}{\sigma}\right)^3\right]$

déjà observer le point (3) des faits stylisés, c'est-à-dire le regroupement des volatilités. En effet, les périodes de forte et de faible volatilité semblent regroupées relativement distinctivement pour cette série. En deuxième lieu, une analyse des statistiques descriptives obtenues pour notre échantillon est présentée dans le tableau 1.1.

Tableau 1.1 – Statistiques descriptives sur le S&P 500

Moyenne	Écart-type	Asymétrie ($\tilde{\mu}_3$)	Aplatissement ($\tilde{\mu}_4$)	Min	Max	n
0.026	1	-1.005	29.89	-22.9	10.957	14882

Ce tableau indique que notre série affiche une asymétrie négative et un aplatissement presque dix fois supérieur à celui d'une loi normale, ce qui concorde avec le point (4).

Maintenant, la figure 1.2 présente les fonctions d'autocorrélation des log-rendements et des log-rendements centrés puis mis au carré (ϵ_t^2) ou en valeur absolue ($|\epsilon_t|$). Il est intéressant de noter que $\mathbb{C}\text{orr}\left[\epsilon_t^2, \epsilon_{t+h}^2\right] < \mathbb{C}\text{orr}[|\epsilon_t|, |\epsilon_{t+h}|]$, $\forall h > 0$. Ce phénomène est appelé l'effet de Taylor [5] [6] [7] et un modèle adéquat devrait, selon Granger et Ding [1995], respecter l'effet de Taylor $\forall h > 0$. Cette figure illustre clairement les points (1) et (2).

La figure 1.3, quant à elle, illustre l'effet de levier présenté au point (5). Le premier graphique de cette figure présente la $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{-\epsilon_t, 0\}]$, le second présente $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{\epsilon_t, 0\}]$ et le troisième la différence entre les deux. Il est clair que notre série confirme la présence de cet effet de levier, puisque $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{-\epsilon_t, 0\}] - \mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{\epsilon_t, 0\}] > 0$, pour presque tout $h \geq 0$.

Remarque 1.1.2. À la lumière de ces caractéristiques, il semble raisonnable de considérer, en vertu des faits stylisés (2) et (3), que l'accent de la modélisation de séries financières soit mis sur la volatilité conditionnelle σ_t. En accord avec le fait stylisé (1) et puisque nous n'observons pas, ou que très peu, d'autocorrélation dans la série des rendements, nous pourrions de surcroît faire l'hypothèse d'unicité de la moyenne ($\mu_t = \mu$, $\forall t$). Cette hypothèse permet entre autres de réduire le nombre de paramètres des modèles.

Il est attendu que les modèles de Markov à variables latentes, présentés au chapitre 3, sauront reproduire la majorité de ces faits stylisés, justifiant ainsi leur utilisation dans le milieu financier.

5. Taylor [1986].
6. Granger et Ding [1995].
7. Rydén, Teräsvirta et Åsbrink [1998].

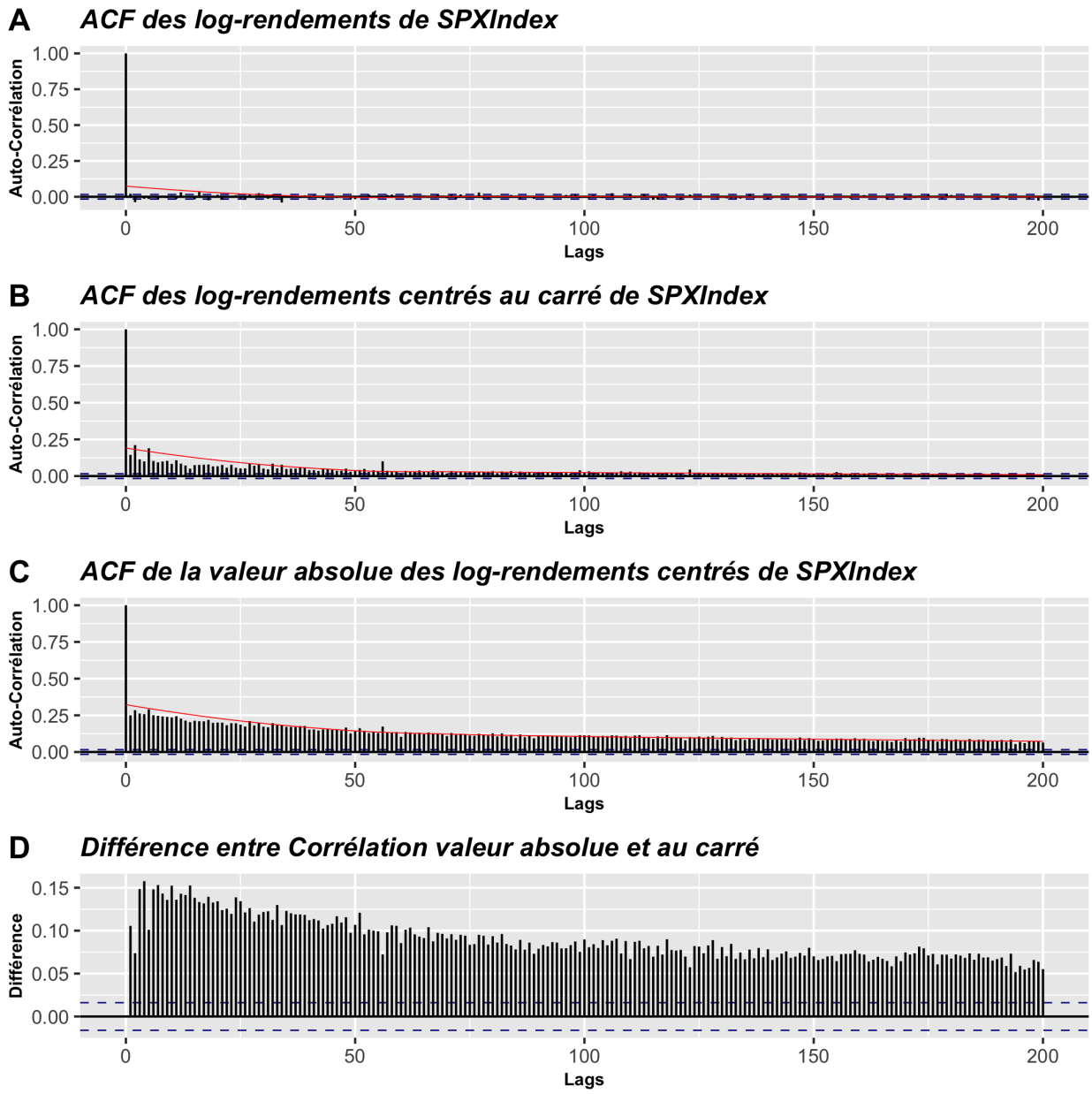


Figure 1.2 – Autocorrélation des log-rendements et des log-rendements centrés en valeur absolue et au carré sur le S&P 500

1.2. Distribution normale univariée

Selon Granger et Ding [1995], il est commun lorsque l'on étudie les rendements sur un actif financier de se concentrer sur deux mesures : une mesure de positionnement et une mesure de dispersion. Puisqu'elle adresse directement ces deux mesures par le biais de ses paramètres, la distribution normale est une candidate de choix pour ce genre d'étude.

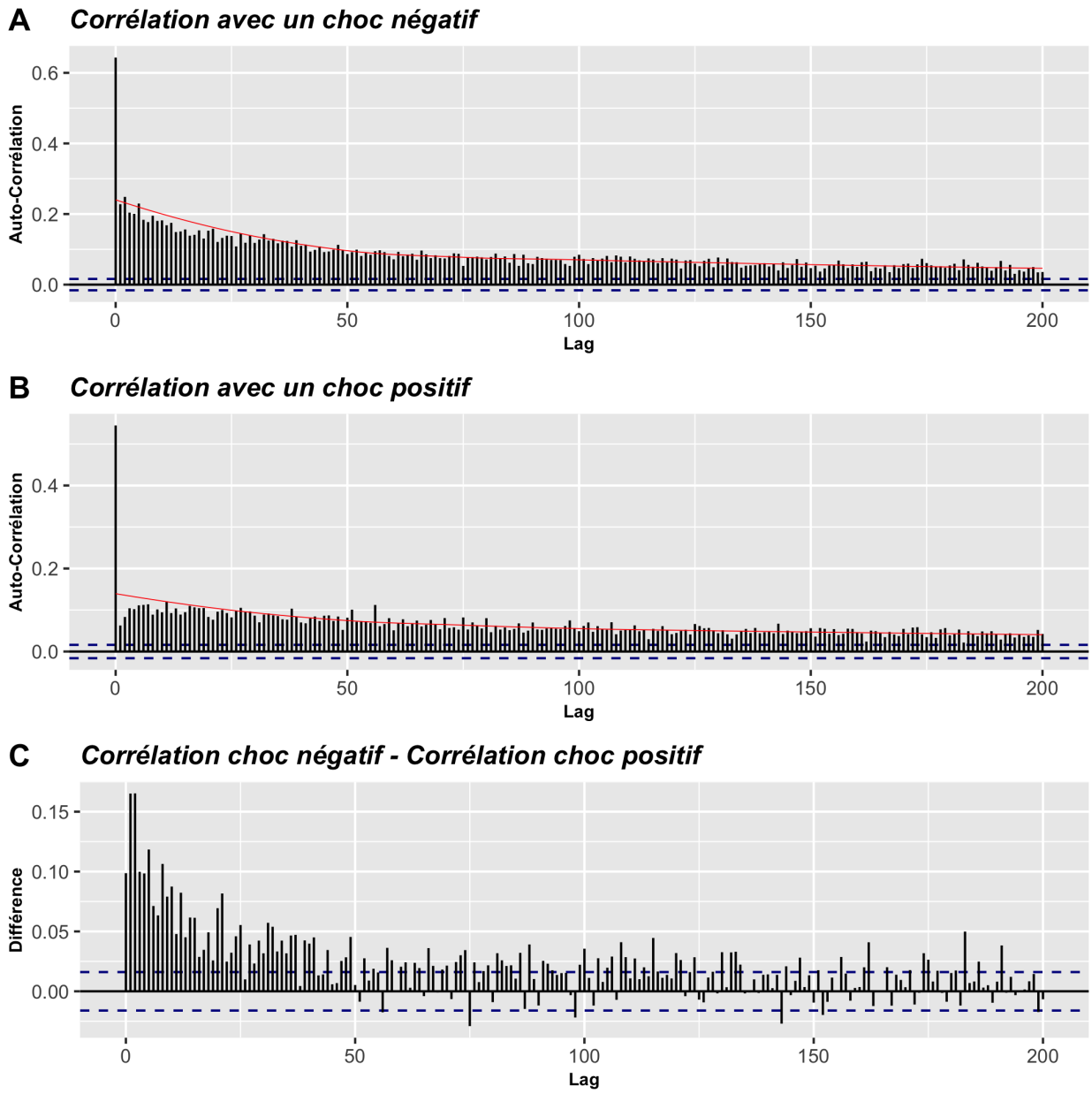


Figure 1.3 – Illustration de l'effet de levier sur les log-rendements du S&P 500

Ainsi, les modèles statistiques présentés dans ce mémoire seront successivement de plus en plus complexes, mais auront toujours en leur coeur des distributions normales univariées. Celles-ci, notées en général $\mathcal{N}(\mu,\sigma^2)$, ont pour fonction de densité

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{-\frac{1}{2\sigma}(x-\mu)^2\Big\}. \tag{1.2.1}$$

De plus, les moments d'une variable aléatoire normale univariée sont obtenus directement, à même les paramètres du modèle

$$\mathbb{E}[X] = \mu \in \mathbb{R}, \tag{1.2.2}$$

$$\mathbb{V}ar[X] = \sigma^2 \in \mathbb{R}_+^\star. \tag{1.2.3}$$

En utilisant d'abord les outils fournis par R, nous testons pour la normalité de l'échantillon.

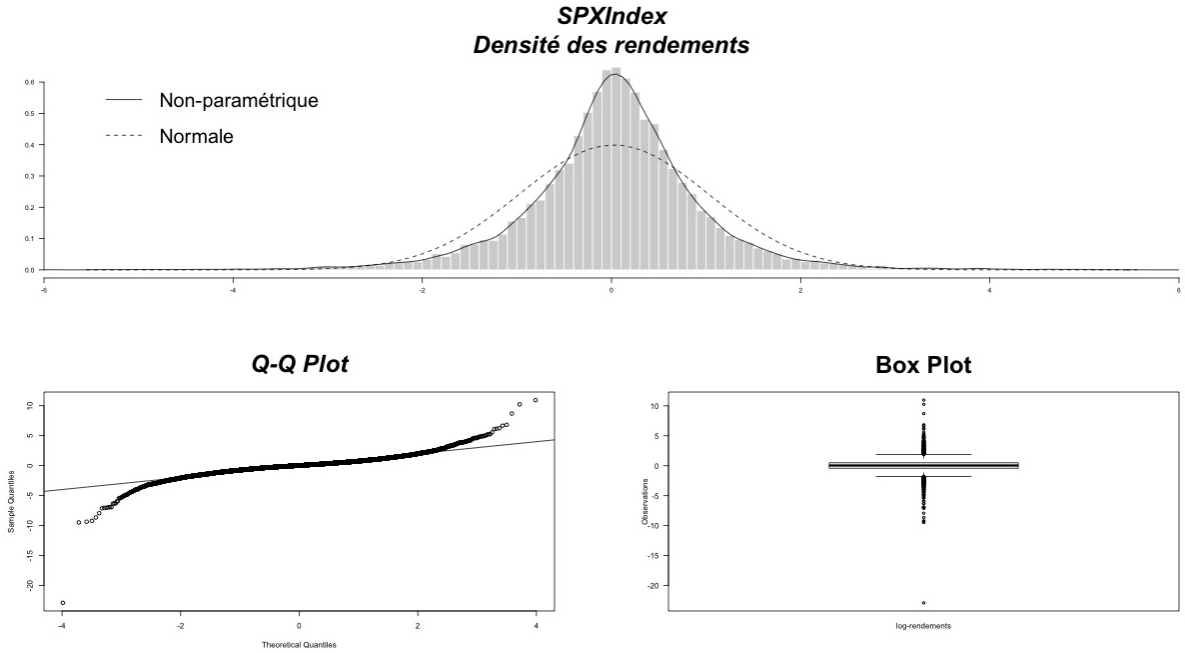


Figure 1.4 – Densité, test des quantiles et *Box Plot* pour les log-rendements du S&P 500 [8]

Le premier graphique de la figure 1.4 semble indiquer une forme presque normale pour la densité non-paramétrique des log-rendements. Il semble par contre que la distribution ait des ailes trop épaisses, ce qui concorde avec le fait stylisé (4). De plus, le second graphique indique une grande dispersion dans les données comparativement à une distribution normale univariée. Cette dispersion est aussi très apparente dans le *Box Plot* sur lequel nous observons une grande quantité de valeurs aberrantes. Une lecture rapide du tableau 1.1 des statistiques descriptives sur le S&P 500 indique aussi une bien plus grand dispersion dans nos données par rapport à un échantillon normalement distribué. Ceci pourrait indiquer que plus d'une distribution est en cause dans notre échantillon. C'est justement cette dispersion qui motive, à ce stade-ci, l'introduction des modèles de mélange.

8. Sur le *Box Plot*, les limites inférieure et supérieure des deux bandes horizontales grises représentent respectivement les quartiles Q_1 et Q_3 (avec $IQR = Q_3 - Q_1$), la ligne les séparant représente la médiane et les deux traits horizontaux représentent le *minimum* ($Q_1 - 1.5IQR$) et le *maximum* ($Q_3 + 1.5IQR$). Les points apparaissant en dessous du *minimum* et au-dessus du *maximum* correspondent à des valeurs aberrantes.

1.3. Modèles de mélange

Lorsque l'on observe un processus et qu'il semble que plus d'une distribution statistique soit en cause, il est commun de considérer l'utilisation de modèles de mélange. Ceux-ci, comme leur nom l'indique, sont des modèles proposant à chaque observation un mélange de distributions statistiques. Ces modèles servent, entre autres, à corriger un problème de sur-dispersion dans les données pouvant réduire énormément la qualité de l'inférence statistique.

Définition 1.3.1 (Processus stochastique)**.** *Un processus stochastique est une famille—ou suite—de variables aléatoires $\{C_t\}_{t\in\mathbb{N}}$, indicées et ordonnées, représentant l'évolution dans le temps d'un phénomène aléatoire.*

Définition 1.3.2 (Modèle de mélange)**.** *Soit K, le nombre de distributions considérées dans le mélange et t, le temps. Un modèle de mélange considère 2 processus stochastiques. Le premier processus est noté $\{C_t\}_{t\in\mathbb{N}^\star}$, est considéré iid et n'est pas observé (latent). La t^e variable de ce processus génère une valeur $i \in \{1, 2, \ldots, K\}$. C'est cette valeur que prend C_t qui régit la distribution du second processus au temps t. Ce second processus, noté $\{X_t\}_{t\in\mathbb{N}^\star}$ est celui qui génère l'observation x_t. Ce type de modèle stipule que la valeur prise par X_t est indépendante de ses valeurs passées et ne dépend que de la valeur prise par C_t. Les propriétés habituelles de ce modèle, $\forall\, k \in \{1, \ldots, K\}$ et $\forall\, t \in \mathbb{N}$, sont résumées à la spécification d'un modèle de mélange* (p.42)*.*

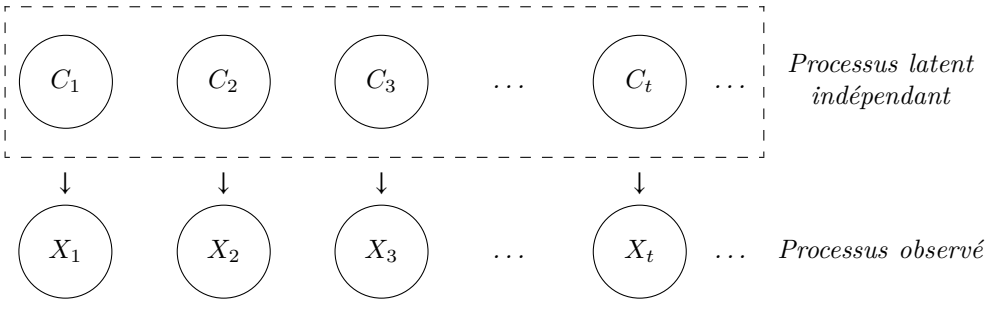
Spécification 1 : Modèle de mélange

Distribution de mélange	$C_t \sim \mathbb{P}[C_t = k] = \pi_k$
Probabilité d'émission	$X_t \mid C_t = k \sim p_k(x_t)$

Nouvelle notation associée

$$p_i(x_t) \stackrel{\text{def}}{=} \mathbb{P}[X_t = x_t \mid C_t = i] \tag{1.3.1}$$

Représentation graphique du modèle complet



La différence majeure entre ce type de modèle et les modèles de Markov à variables latentes qui seront introduits à la section 3 est que l'on considère dans le cas présent que le processus $\{C_t\}_{t\in\mathbb{N}^\star}$ est lui aussi *iid*.

Les distributions marginales univariées du processus $\{X_t\}_{t\in\mathbb{N}^\star}$ sont obtenues en conditionnant sur les valeurs du processus $\{C_t\}_{t\in\mathbb{N}^\star}$ et par le biais de la formule des probabilités totales. Dans un modèle de mélange à K composantes, les distributions marginales des variables aléatoires X_t sont donc obtenues ainsi

$$\begin{aligned}\mathbb{P}[X_t = x_t] &= \sum_{i=1}^{K} \mathbb{P}[X_t = x_t \mid C_t = i]\mathbb{P}[C_t = i], \\ &= \sum_{i=1}^{K} \pi_i p_i(x_t), \end{aligned} \tag{1.3.2}$$

où $p_i(x_t)$ pourrait être une fonction de densité ou de masse.

En représentant les distributions marginales des variables aléatoires C_t sous la forme de vecteurs lignes de dimension K,

$$\vec{\pi} = (\mathbb{P}[C_t = 1], \ldots, \mathbb{P}[C_t = K]), \quad \forall\, t \in \mathbb{N}^\star, \tag{1.3.3}$$

$$= (\pi_1, \ldots, \pi_K), \tag{1.3.4}$$

cette équation peut s'écrire sous la forme matricielle

$$\mathbb{P}[X_t = x_t] = (\pi_1, \ldots, \pi_K) \begin{bmatrix} p_1(x_t) & & 0 \\ & \ddots & \\ 0 & & p_K(x_t) \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{1.3.5}$$

$$= \vec{\pi}\ \mathbf{P}(x_t)\mathbf{1}_K, \tag{1.3.6}$$

avec $\mathbf{P}(x_t) = \text{diag}(\vec{p}(x_t)) = \text{diag}(p_1(x_t), \ldots, p_K(x_t))$ et où $\mathbf{1}_K = (1 \cdots 1)'$ est un vecteur colonne de dimension K dont toutes les entrées valent 1.

Les distributions marginales bivariées sont obtenues similairement,

$$\mathbb{P}[X_t = x_t, X_{t+k} = x_{t+k}] = \sum_{i,j=1}^{K} \mathbb{P}[X_t = x_t, X_{t+k} = x_{t+k}, C_t = i, C_{t+k} = j], \tag{1.3.7}$$

$$= \sum_{i,j=1}^{K} \mathbb{P}[X_t = x_t \mid C_t = i]\mathbb{P}[C_t = i]\mathbb{P}[X_{t+k} = x_{t+k} \mid C_{t+k} = j]\mathbb{P}[C_{t+k} = j],$$

$$= (\vec{\pi}\ \mathbf{P}(x_t)\mathbf{1}_K) \times (\vec{\pi}\ \mathbf{P}(x_{t+k})\mathbf{1}_K). \tag{1.3.8}$$

Les moments de X_t s'obtiennent aussi simplement. En suivant la même logique, nous avons que l'espérance d'une fonction $g(x)$ de X_t est

$$\begin{aligned}\mathbb{E}[g(X_t)] &= \sum_{i=1}^{K} \mathbb{E}[g(X_t) \mid C_t = i]\mathbb{P}[C_t = i], \\ &= \sum_{i=1}^{K} \pi_i \mathbb{E}[g(X_t) \mid C_t = i], \end{aligned} \tag{1.3.9}$$

lorsque l'espérance est définie pour la transformation $g(X_t)$ et où $\mathbb{E}[g(X_t) \mid C_t = i]$ dépend de la distribution associée à la i^{e} composante du modèle.

En notant $\mu_{g|i} \overset{\text{def}}{=} \mathbb{E}[g(X_t) \mid C_t = i]$, $\boldsymbol{\mu}_g \overset{\text{def}}{=} \big(\mu_{g|1}, \ldots, \mu_{g|K}\big)'$ et $\mathbf{M}_g \overset{\text{def}}{=} \text{diag}\big(\boldsymbol{\mu}_g\big)$, nous pouvons réécrire cette somme sous la forme matricielle

$$\mathbb{E}[g(X_t)] = \vec{\pi}\mathbf{M}_g \mathbf{1}_K. \tag{1.3.10}$$

1.3.1. Estimation des paramètres

Puisque l'une des probabilités du mélange peut être obtenue par soustraction $\left(\sum_{i=1}^{K} \pi_i = 1 \equiv \pi_K = 1 - \sum_{i=1}^{K-1} \pi_i\right)$, les paramètres d'un modèle à K composantes se résument à $(K-1)$ probabilités pour le mélange et K vecteurs de paramètres pour les distributions du mélange. Pour un mélange de lois normales $\mathcal{N}(\mu_i,\sigma_i^2)$, ceci revient donc à $3K-1$ paramètres, contenus dans un vecteur noté $\boldsymbol{\theta}$. Les paramètres optimaux pour un échantillon donné sont obtenus en maximisant la distribution conjointe de toutes les observations, aussi appelée vraisemblance du modèle. Les paramètres ainsi obtenus sont appelés estimateurs du maximum de vraisemblance (EMV) et nous noterons $\widehat{\boldsymbol{\theta}}$ le vecteur contenant ces paramètres estimés.

Plusieurs notations équivalentes existent pour identifier la vraisemblance (ou distribution marginale de grand ordre) d'un échantillon, quelques unes de ces notations sont répertoriées dans le tableau 1.2.

$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T]$		
$f(x_{1:T} \mid \boldsymbol{\theta})$,	$\mathbb{P}[X_{1:T} = x_{1:T}]$,	$\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})$,
$\mathcal{L}_T(\boldsymbol{\theta})$,	$\mathcal{L}(\boldsymbol{\theta} \mid x_{1:T})$,	$\mathcal{L}_T$.

Tableau 1.2 – Diverses notations pour la vraisemblance d'un échantillon $x_{1:T}$

Par soucis de précision et de concision, nous retiendrons une version légèrement alourdie de la notation utilisée dans Zucchini et Macdonald [2009](p. 37), soit $\mathcal{L}_T(\boldsymbol{\theta})$.

Remarque 1.3.1. Lorsqu'il est question de probabilités—ou de fonctions de celles-ci—il serait plus juste d'inscrire explicitement le vecteur de paramètres sur lequel la probabilité est conditionnelle. Il aurait donc, par exemple, été plus juste d'écrire à la première notation du tableau 1.2

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T \mid \boldsymbol{\theta}]. \tag{1.3.11}$$

Cependant, cette écriture est plus lourde et c'est surtout lorsqu'il est question d'estimation de vecteurs de paramètres qu'il devient important de noter à quels paramètres la fonction de probabilité fait référence. C'est pourquoi ceux-ci seront seulement spécifiés dans le symbole retenu pour la vraisemblance, $\mathcal{L}_T(\boldsymbol{\theta})$.

Pour un échantillon de T observations, avec un total de p paramètres pour les distributions et K distributions, cette vraisemblance s'écrit donc

$$\begin{aligned}
\mathcal{L}_T(\boldsymbol{\theta}) &\overset{\text{def}}{=} \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_T], && (1.3.12)\\
&\overset{ind.}{=} \prod_{t=1}^{T} \mathbb{P}[X_t = x_t], \\
&= \prod_{t=1}^{T} \sum_{i=1}^{K} \pi_i p_i(x_t), \\
&= \prod_{t=1}^{T} \sum_{i=1}^{K} \pi_i \left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\,\sigma_i} (x_t - \mu_i)^2 \right\} \right), && (1.3.13)
\end{aligned}$$

où $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p, \pi_1, \ldots, \pi_{K-1})' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{(p+K)}$ et $\boldsymbol{\Theta}$ représente l'espace des paramètres.

Ainsi, en termes mathématiques, les EMVs satisfont la relation suivante :

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\, x_{1:T} \mid \boldsymbol{\theta}\,), \tag{1.3.14}$$

ou de manière équivalente

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_T(\boldsymbol{\theta}). \tag{1.3.15}$$

La recherche du maximum d'un produit n'est cependant pas chose aisée. Mis à part certains algorithmes comme le *MO-CMA-ES*[9] présenté par Langouët [2011] et ayant recours à des interpolations, la maximisation numérique d'une fonction passe presque toujours par le calcul de son gradient. L'objectif est donc en partie de trouver une forme pour la vraisemblance dont la dérivation est relativement simple à obtenir.

Puisque le logarithme d'un produit se simplifie à la somme du logarithme des termes séparés, il est bien plus pratique de chercher le maximum du logarithme de la vraisemblance. Dans ce cas, le gradient peut être calculé terme par terme. De plus, cette transformation

9. *Multi-Objective - Covariance Matrix Adaptation - Evolution Strategy*

est justifée mathématiquement puisque la fonction logarithmique est une fonction monotone croissante, conservant donc la position du maximum. La log-vraisemblance du même modèle s'écrit

$$\ell_T(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log(\mathcal{L}_T(\boldsymbol{\theta})), \tag{1.3.16}$$

$$= \sum_{t=1}^{T} \log\left(\sum_{i=1}^{K} \pi_i \left(\frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_i}(x_t - \mu_i)^2\right\}\right)\right). \tag{1.3.17}$$

Cette fonction ne peut malheureusement pas non plus être optimisée analytiquement, il est donc nécessaire d'avoir recours à des routines d'optimisation numérique. Celles-ci sont incluses dans les langages de programmation comme `R`, qui offre par exemple les fonctions `nlm` et `optim` permettant de trouver le minimum d'une fonction donnée. On cherchera donc plutôt le minimum de $-\ell_T(\boldsymbol{\theta})$ afin de trouver les paramètres optimaux du modèle.

Les modèles de mélange sont forts utiles afin d'étendre les ailes d'une distribution et régler ainsi le problème de surdispersion. Cependant, rappelons qu'à la section 1.1 nous affirmions que les rendements financiers affichent habituellement une forme de dépendance temporelle. Évidemment, cette hypothèse de dépendance est antithétique à l'hypothèse d'indépendance du processus $\{C_t\}_{t\in\mathbb{N}^\star}$ des modèles de mélange.

Une approche classique consiste à intégrer cette dépendance à même le processus des rendements. Ce type d'approche rassemble entre autres les modèles de type autorégressif (AR) et à moyenne mobile (MA)—ayant tous deux pour but de générer une dépendance linéaire dans le processus des rendements—et autorégressif-conditionnellement-hétéroscédastique ($ARCH$)—générant plutôt une dépendance dans le processus de la valeur absolue ou du carré des rendements. Puisqu'en vertu du fait stylisé (1) (p.36), nous n'observons pas de dépendance linéaire dans le processus de rendement, ce n'est pas une dynamique qui nécessite d'être incluse dans le modèle financier. Ainsi, les modèles de type $ARCH$ seront habituellement préférés aux modèles de types AR et MA pour l'étude de séries financières.

Une autre approche consiste à modéliser cette dépendance par le biais d'un processus latent régissant le processus des rendements. Cette approche peut être vue comme une généralisation naturelle des modèles de mélange, ce qui la rend spécialement attrayante.

Tel que démontré notamment dans Augustyniak [2014], ces deux approches peuvent aisément être combinées. Nous choisissons cependant de concentrer nos efforts sur la seconde approche dans les prochains chapitres.

Chapitre 2

Chaîne de Markov à temps discret

Dans ce chapitre, ainsi que dans les chapitres subséquents, il sera question de matrices carrées non-négatives $\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j\in\{1,\dots,K\}}$, dont toutes les entrées satisfont $\gamma_{ij} \geq 0$. Nous utiliserons alors la notation $\boldsymbol{\Gamma} \geq 0$ pour signifier cette propriété et $\boldsymbol{\Gamma} > 0$ lorsque $\gamma_{ij} > 0$, $\forall i, j$. Nous écrirons donc que $\boldsymbol{\Gamma} \geq \mathbf{B} \equiv \boldsymbol{\Gamma} - \mathbf{B} \geq 0$, avec des matrices $\boldsymbol{\Gamma}$ et $\mathbf{B}$ de dimensions cohérentes. Cette notation s'étend naturellement aux vecteurs.

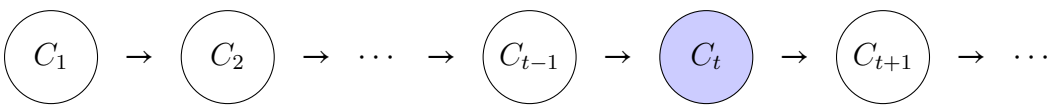
L'objet de cet ouvrage est principalement les modèles de Markov à variables latentes. À cet effet, il convient au préalable de nous intéresser aux chaînes de Markov et en particulier à celles à temps discret et à espace d'états fini. Ces chaînes étant au coeur même des modèles en question, nous débuterons par une présentation de quelques définitions et propriétés de celles-ci, nous permettant ensuite d'intégrer ces caractéristiques à la description des modèles de Markov à variables latentes présentés au chapitre 3 (p.53).

2.1. Définitions

Définition 2.1.1 (Chaîne de Markov)**.** *Une chaîne de Markov à temps discret et à espace d'états dénombrable est un processus stochastique, prenant des valeurs dans un espace d'états dénombrable $\mathcal{C}$ et possédant la propriété markovienne. Cette propriété stipule que toute l'information sur la valeur suivante de la chaîne se trouve dans sa valeur présente et ne dépend pas des valeurs antérieures à cette dernière. Cette propriété de « perte de mémoire » se résume à la simplification suivante :*

$$\mathbb{P}[C_{t+1} = c_{t+1} \mid C_{1:t}] = \mathbb{P}[C_{t+1} = c_{t+1} \mid C_t], \quad \forall\, t \in \mathbb{N}, \tag{2.1.1}$$

et graphiquement, la chaîne peut être vue ainsi



En particulier, nous nous concentrerons sur les chaînes de Markov à temps discret et à espace d'états fini $\mathcal{C} \in \mathbb{R}^K$. Ceci permettra, sans perte de généralité, d'indicer les valeurs possibles de la variable aléatoire et de définir l'espace d'états $\mathcal{C} = \{1, 2, \ldots, K\}$. Ainsi, $\forall\, i, j \in \{1, \ldots, K\}$ et $\forall\, t \in \mathbb{N}^\star$, nous noterons la ***probabilité de transition*** *au temps t, de l'état i vers l'état j :*

$$\gamma_{ij}(t) \stackrel{\text{def}}{=} \mathbb{P}[C_{t+1} = j \mid C_t = i]. \tag{2.1.2}$$

Définition 2.1.2 (Matrice de transition)**.** *Les probabilités de transition d'une chaîne de Markov sont rassemblées dans une matrice appelée* ***matrice de transition****. Ces probabilités sont inscrites dans cette matrice de sorte que les lignes correspondent à l'état présent et les colonnes à l'état suivant de la chaîne. Ainsi, la probabilité $\gamma_{ij}, i, j \in \mathcal{C}$, se retrouve à l'intersection de la i^e ligne et de la j^e colonne de cette matrice carrée, que nous noterons en général $\mathbf{\Gamma}_t$:*

$$\mathbf{\Gamma}_t = \begin{bmatrix} \gamma_{11}(t) & \cdots & \gamma_{1K}(t) \\ \vdots & \gamma_{ij}(t) & \vdots \\ \gamma_{K1}(t) & \cdots & \gamma_{KK}(t) \end{bmatrix}, \tag{2.1.3}$$

avec comme condition que

$$\sum_{j=1}^{K} \gamma_{ij} = 1, \quad \forall\, i \in \mathcal{C}, \tag{2.1.4}$$

ou de manière équivalente

$$\mathbf{\Gamma}_t \mathbf{1}_K = \mathbf{1}_K. \tag{2.1.5}$$

Cette condition, ainsi que le fait que $\mathbf{\Gamma}_t \geq 0$, qualifie la matrice de transition d'une chaîne de Markov de matrice stochastique.

Définition 2.1.3 (Homogénéité)**.** *Lorsque les probabilités de transition ne dépendent pas de l'indice t, nous dirons que celles-ci sont* ***homogènes****. Dans ce cas, on dit de la chaîne qu'elle est homogène et la probabilité de transition d'un état $C_t = i$ vers un état $C_{t+1} = j$, avec i,$j \in \mathcal{C}$, est alors notée*

$$\gamma_{ij} \stackrel{\text{def}}{=} \mathbb{P}[C_{t+1} = j \mid C_t = i], \quad \forall t. \tag{2.1.6}$$

De plus, lorsque la chaîne est homogène, la matrice de transition est unique et notée

$$\mathbf{\Gamma} = (\gamma_{ij})_{i,j \in \mathcal{C}}. \tag{2.1.7}$$

À moins que ce ne soit indiqué par l'absence de l'indice t ou précisé autrement, nous ferons l'hypothèse de non-homogénéité et c'est donc la notation $\mathbf{\Gamma}_t$ qui sera privilégiée pour les définitions suivantes. Certaines architectures des modèles de Markov à variables latentes sont construites avec une chaîne de Markov non-homogène par définition. Quelques unes de ces architectures seront discutées au chapitre 5.

2.2. Matrice de transition en m pas

La matrice de transition telle que définie précédemment permet d'avancer d'un pas dans la chaîne de Markov. Lorsque l'on cherche à avancer de $m \in \mathbb{N}^\star$ pas en avant, nous utiliserons la matrice de transition en m pas que nous noterons simplement $\mathbf{\Gamma}_t^{(m)}$. Cette matrice contient, au croisement de la i^{e} ligne et de la j^{e} colonne, la probabilité

$$\gamma_{ij}^{(m)}(t) = \mathbb{P}[C_{t+m} = j \mid C_t = i]. \tag{2.2.1}$$

Remarque 2.2.1. Lorsque la chaîne est homogène, ces probabilités sont notées

$$\gamma_{ij}^{(m)} = \mathbb{P}[C_{t+m} = j \mid C_t = i], \forall\, t \in \mathbb{N}^\star. \tag{2.2.2}$$

Quelques propriétés supplémentaires des matrices de transition en m pas et des probabilités de transition sont reportées dans l'appendice externe.

2.3. Classification des états

Les propriétés des états d'une chaîne de Markov sont elles aussi reportées dans l'appendice externe. La propriété la plus importante, l'ergodicité, est présentée ici parce qu'elle est centrale à la théorie des chaînes de Markov. Nous introduisons de plus les propriétés d'irréductibilité et de périodicité puisqu'elles sont cruciales à la compréhension de l'ergodicité.

Définition 2.3.1 (Communication entre les états)**.** *Un état j est dit accessible à partir d'un état i, noté $i \longrightarrow j$, si et seulement si $\exists\ m(i,j) \in \mathbb{N}^\star$ t.q. $\gamma_{ij}^{(m(i,j))} > 0$. Si, de plus, i est accessible à partir de j, nous dirons que les deux états communiquent et cette relation sera notée $i \longleftrightarrow j$.*

Si tous les états de la chaîne communiquent entre eux, ils sont dits faire partie de la même classe d'équivalence. Lorsque l'espace d'états est composé d'une seule classe d'équivalence, la chaîne est dite ***irréductible****.*

Définition 2.3.2 (Période)**.** *La période d'un état k d'une chaîne de Markov est définie comme*

$$d(k) \stackrel{\text{def}}{=} \text{pgcd}\Big(n \geq 1 : \gamma_{kk}^{(n)} > 0\Big). \tag{2.3.1}$$

Un état k est alors dit ***apériodique*** *si $d(k) \in \{0, 1\}$ et* ***périodique*** *lorsque $d(k) > 1$. Par convention, $d(k) = 0$ si l'état k n'est pas accessible à partir de lui-même.*

Définition 2.3.3 (Ergodicité)**.** *Une chaîne de Markov irréductible, apériodique et sur un espace d'états fini est dite* ***ergodique****.*

2.4. Distribution stationnaire

Si une chaîne de Markov est ergodique et homogène, elle admet une distribution stationnaire unique notée $\vec{\pi} = (\pi_1, \ldots, \pi_K) > 0$, telle que

$$\lim_{m\to\infty} \mathbf{\Gamma}^{(m)} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_K \\ \pi_1 & \pi_2 & \cdots & \pi_K \\ & & & \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_K \end{bmatrix}. \tag{2.4.1}$$

Bien entendu, puisque le vecteur $\vec{\pi}$ correspond à une ligne de la limite de la matrice de transition en m pas, elle est soumise à toutes les contraintes auxquelles sont soumises les lignes de la matrice $\mathbf{\Gamma}$ d'une chaîne ergodique. En plus de ces contraintes cependant, puisqu'une chaîne ergodique est par définition irréductible, les termes de la distribution stationnaire doivent être tous non-nuls.

De plus, cette distribution stationnaire satisfait

$$\vec{\pi} = \vec{\pi}\mathbf{\Gamma} \tag{2.4.2}$$

ou alternativement

$$\pi_i = \sum_{j=1}^{K} \pi_j \gamma_{ji}, \quad \forall\, i \in \mathcal{C}. \tag{2.4.3}$$

C'est généralement en résolvant l'équation (2.4.2) que la distribution stationnaire de la chaîne de Markov est obtenue.

Remarque 2.4.1 (Irréductibilité et identifiabilité). Ici nous bornons π_i strictement, tel que les cas $\pi_i = 0$ et $\pi_i = 1$ soient exclus. Ceci est une conséquence du théorème ergodique qui

nécessite l'irréductibilité de la chaîne. De plus, ceci est lié au concept d'identifiabilité d'un modèle de Markov à variables latentes qui est discuté dans l'appendice externe.

Afin d'alléger le texte, les définitions autour de la diagonalisation de la matrice de transition ainsi que le théorème de Perron-Frobenius sont aussi reportés dans l'appendice externe. Il est toutefois important de mentionner ici que c'est grâce à ce théorème que l'on peut affirmer l'existence et l'unicité de la distribution stationnaire d'une chaîne de Markov ergodique.

Une autre conséquence importante de ce théorème est que la plus grande valeur propre de la matrice $\mathbf{\Gamma}$ est $\lambda_1 = 1$ et que la seconde valeur propre $\lambda_2 < 1$ indique la vitesse de convergence géométrique de la matrice $\mathbf{\Gamma}$ vers sa distribution stationnaire, donnant une idée de la persistance combinée des états de la chaîne. Lorsque les régimes sont persistants, ces modèles permettent habituellement de reproduire le regroupement des volatilités, soit le fait stylisé (3) (p.36).

2.5. Distributions marginales

Comme à la section 1.3, les distributions marginales des variables aléatoires C_t seront représentées sous la forme de vecteurs lignes de dimension K. Cependant, dans le cas des *HMMs*, celles-ci pourraient changer à chaque pas et dépendent donc de l'indice t. Nous noterons ces probabilités

$$\vec{\mathfrak{u}}(t) = (\mathbb{P}[C_t = 1], \ldots, \mathbb{P}[C_t = K]), \quad \forall\, t \in \mathbb{N}. \tag{2.5.1}$$

$$= (\mathfrak{u}_1(t), \ldots, \mathfrak{u}_K(t)) \tag{2.5.2}$$

avec $\mathfrak{u}_k(t) = \mathbb{P}[C_t = k]$.

En particulier, lorsque la chaîne de Markov débute en $t = 1$, la distribution initiale de la chaîne est notée $\vec{\mathfrak{u}}(1)$. Celle-ci est cruciale à la caractérisation d'une chaîne de Markov, en particulier pour l'obtention des distributions conjointes. Nous référons les intéressé.e.s à l'appendice externe pour plus de détails.

Les distributions marginales suivantes sont obtenues récursivement à partir de la distribution initiale de la chaîne ainsi que la matrice de transition. En effet, nous pouvons obtenir directement la distribution marginale de la chaîne à n'importe quel point $(t + h)$ dans le temps à partir de la distribution marginale à t et la matrice de transition :

$$\vec{\mathfrak{u}}(t+h) = \vec{\mathfrak{u}}(t)\mathbf{\Gamma}_t^{(h)}, \qquad \forall\, t,h \in \mathbb{N}^\star, \tag{2.5.3}$$

$$= \vec{\mathfrak{u}}(t)\mathbf{\Gamma}^h, \qquad \text{si la chaîne est homogène.} \tag{2.5.4}$$

Démonstration. Démonstration dans l'appendice externe. □

Remarque 2.5.1. Dans le cas où la chaîne est homogène, soit l'équation (2.5.4), la notation $\mathbf{\Gamma}^h$ correspond à un produit matriciel. En contrepartie, la notation $\mathbf{\Gamma}_t^{(h)}$ de l'équation (2.5.3) ne fait que préciser que le nombre de pas à prendre à partir de t correspond à h, masquant ainsi les transitions intermédiraires par soucis de concision. Le cas où la chaîne est homogène n'est qu'un cas particulier du cas non-homogène, puisque nous avons toujours que

$$\mathbf{\Gamma}_t^{(h)} = \prod_{\ell=1}^{h} \mathbf{\Gamma}_{t+\ell-1}. \tag{2.5.5}$$

Maintenant que nous avons ces concepts en main, il est temps de combiner les modèles de mélange et les chaînes de Markov pour obtenir les modèles de Markov à variables latentes.

Chapitre 3

Modèles de Markov à variables latentes

Les modèles de Markov à variables latentes ont fait leurs preuves maintes et maintes fois depuis le début de leur étude par le mathématicien américain Leonard E. Baum [1966 ; 1967 ; 1968 ; 1970 ; 1972]. Il désirait alors pouvoir appliquer ces modèles aux marchés boursiers, à l'étude des taches solaires et à la météorologie. C'est d'ailleurs durant ces travaux qu'il développe l'algorithme de Baum-Welch, une méthode itérative permettant d'obtenir les paramètres d'un modèle de Markov à variables latentes, même lorsqu'une forme analytique pour la vraisemblance est très difficile à obtenir. Cet algorithme est d'ailleurs le précurseur de l'algorithme EM, une version aux applications plus générales développée par Dempster, Laird et Rubin [1977]. En présentant une méthode permettant d'estimer les paramètres, il lançait une invitation à tester ces modèles et à bâtir sur ses travaux afin de populariser la nouvelle classe de modèles qu'il proposait.

Ce n'est certainement pas un hasard si ces modèles sont depuis devenus si populaires. Le concept même de changement de régime dans une série chronologique est à la fois intuitif et naturellement observable. Prenons en exemple la température d'une journée à la suivante. Dans certains pays—à commencer par le Canada—, il est clair que la température d'une journée ne dépend pas seulement de la température de la veille, mais aussi de la saison dans laquelle elle est mesurée. Certaines saisons affichent en effet des variations de températures beaucoup plus volatiles que d'autres. Naturellement, ces changements de saisons ne sont pas fruits du hasard, ni d'une structure absolue, ils sont plutôt la conséquence d'un processus que nous n'observons pas et qui régit leur alternance au gré d'un cycle, lui aussi latent. Comme nous avons pu le constater à la section 1.1, il en va de même pour les rendements financiers et les cycles—saisons—économiques associés.

En effet, il est clair depuis les travaux de Hamilton [1989] que les cycles économiques ont une grande influence sur la volatilité du prix d'un titre ou du niveau d'un indice. Les régimes d'expansion et de récession identifiés naturellement par ses modèles correspondaient à des moments de changements dans les politiques monétaires. Ces observations furent depuis

confirmées à maintes reprises. Par exemple dans les travaux d'Andrew Ang, comme dans Ang et Timmermann [2012], où il est question en détails des prouesses des modèles à changement de régime pour l'analyse du rendement d'actifs et la reproduction des *faits stylisés* en finance.

L'intérêt de ces modèles pour l'analyse spécifique de la volatilité des marchés et des rendements est aussi apparent. Cette idée est discutée entre autres dans les travaux de Rossi et Gallo [2002] et de Liu, Margaritis et Wang [2012], et c'est un sujet presque incontournable lorsqu'il est question des applications des modèles de Markov à variables latentes en finance.

Dans un contexte multivarié, ces modèles permettent aussi que les corrélations entre plusieurs actifs changent d'un régime à un autre, relaxation potentiellement majeure discutée par Ang et Chen [2002]. Au moment de construire un portefeuille, la diversification peut être un enjeux majeur nécessitant ensuite un suivi serré afin que celui-ci reste diversifié. En effet, lorsque les corrélations entre les actifs changent, un portefeuille que l'on croyait diversifié peut ne plus l'être. Cette caractéristique des modèles de Markov les rend donc très attrayant lorsqu'il est question d'allocation d'actifs. Ainsi, cette classe de modèles est utilisée entre autres pour l'allocation dynamique d'actifs telle que discutée par Ang et Bekaert [2002a ; 2002b ; 2004], Bulla et al. [2011], Kritzman, Page et Turkington [2012] et dans les travaux de thèse de Peter Nystrup (e.g. [2015 ; 2017]).

Dans le contexte d'analyse de signal, ces modèles sont en fait si généraux et naturels, qu'un monde sépare parfois les domaines où ils sont utilisés. Que ce soit pour la reconnaissance vocale (e.g. Rabiner [1989], Juang et Rabiner [1991]), la prédiction et l'analyse des précipitations au Brésil (e.g. Robertson, Kirshner et Smyth [2003]), l'analyse de séquences biologiques comme le génôme (e.g. Yoon [2009]), la prévision et l'étude de catastrophes naturelles comme les tremblements de terre (e.g. Zucchini et Macdonald [2009]) et bien d'avantage, ces modèles sont maintenant incontournables. Avec la popularisation de l'apprentissage automatique, ces modèles ont aussi reçu une notoriété supplémentaire. En effet, ils apparaissent dans les travaux de Yoshua Bengio (e.g. Bengio et Frasconi [1994], Bengio [1999]) qui propose leur utilisation dans divers contextes, comme la reconnaissance du langage écrit.

Remarque 3.0.1. Il est important de noter que cette classe de modèles a reçu plusieurs dénominations au fil du temps, le nom de modèle de Markov à variables latentes est bien loin d'être le seul qui soit utilisé, même en français. Quelques exemples communs sont : Modèle à chaîne de Markov cachée, Modèle de Markov caché, Modèle à changement de régime Markovien, automate de Markov à états cachés, ou en anglais : *Hidden Markov model*, *Markov-switching model*, *hidden Markov process*[1], *Markov-dependent mixture*[2], *model subject*

1. Ephraim et Merhav [2002].
2. Le, Leroux, Puterman et Albert [1992].

to Markov regime, *Markov mixture model*. Le nom utilisé dépend naturellement du domaine d'activité de la personne qui l'utilise.

Nous utiliserons « modèle de Markov à variables latentes » lorsque nous ferons référence au modèle par son nom complet, mais l'acronyme *HMM*, de l'anglais *Hidden Markov Model*, sera parfois utilisé afin d'alléger le texte.

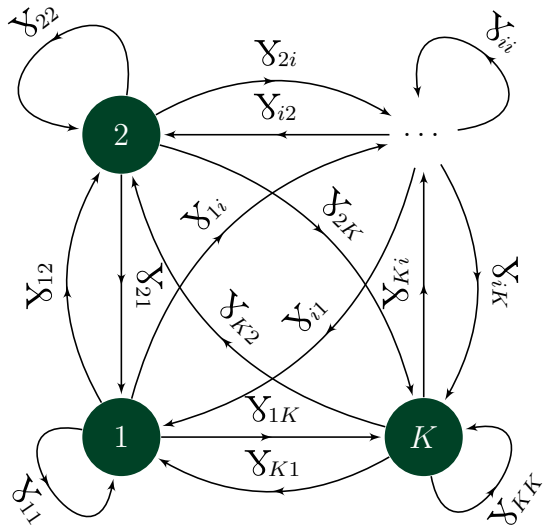
3.1. Définition

Un modèle de Markov à variables latentes est un processus stochastique bivarié à temps discret, noté $\{(X_t, C_t)\}_{t \in \mathbb{N}^\star}$. Les propriétés habituelles de ce modèle sont résumées à la spécification standard (p.56).

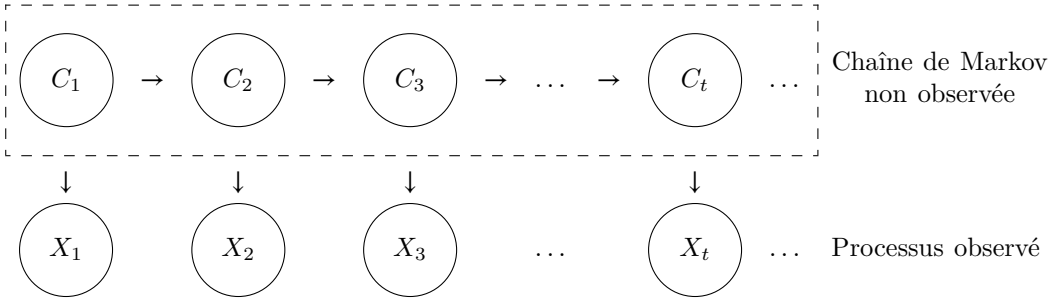
Spécification 2 : HMM standard

Distribution initiale	$C_1 \sim p(c_1)$
Probabilité de transition	$C_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t-1} = c_{1:t-1}) \sim p(c_t \mid c_{t-1})$
Probabilité d'émission	$X_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t} = c_{1:t}) \sim p(x_t \mid c_t)$

Représentation graphique du processus latent



Représentation graphique du modèle complet



Constitués de deux parties, une chaîne de Markov non-observée (ou latente) notée $\{C_t\}_{t\in\mathbb{N}^\star}$ et un processus observé noté $\{X_t\}_{t\in\mathbb{N}^\star}$, ce modèle est tel que lorsque la valeur de C_t est connue, la distribution de X_t ne dépend que de cette valeur. Ceci signifie que, conditionnellement aux valeurs prisent par $\{C_t\}_{t\in\mathbb{N}^\star}$, le processus $\{X_t\}_{t\in\mathbb{N}^\star}$ est une suite de variables aléatoires indépendantes, et donc que

$$\mathbb{P}[x_{1:n} \mid c_{1:T}] = \prod_{t=1}^{n} \mathbb{P}[x_t \mid c_t], \quad n = 1, \ldots, T. \tag{3.1.1}$$

La structure de ces modèles est illustrée à la portion diagramme de la spécification standard (p.56) par un graphique orienté dans lequel la zone délimitée par une ligne discontinue indique que nous n'observons pas son contenu.

Remarque 3.1.1 (Propriété markovienne). Bien que le processus $\{X_t\}_{t\in\mathbb{N}^\star}$ ne soit pas une chaîne de Markov, le processus bivarié $\{(X_t, C_t)\}_{t\in\mathbb{N}^\star}$ possède malgré tout la propriété markovienne de perte de mémoire. En effet, nous avons que

$$\begin{aligned}\mathbb{P}[x_t, c_t \mid x_{1:t-1}, c_{1:t-1}] &\overset{\text{Bayes}}{=\!=} \mathbb{P}[c_t \mid x_{1:t-1}, c_{1:t-1}] \cdot \mathbb{P}[x_t \mid x_{1:t-1}, c_{1:t}],\\ &= \mathbb{P}[c_t \mid x_{t-1}, c_{t-1}] \cdot \mathbb{P}[x_t \mid x_{t-1}, c_{t-1}, c_t],\\ &\overset{\text{Bayes}}{=\!=} \mathbb{P}[x_t, c_t \mid x_{t-1}, c_{t-1}], \end{aligned} \tag{3.1.2}$$

et donc que ce processus bivarié est une chaîne de Markov selon la définition 2.1.1 (p.47).

La définition d'un modèle de Markov à variables latentes standard n'est en fait pas très différente de celle d'un modèle de mélange indépendant (définition 1.3.2 (p.42)). Par exemple, si toutes les lignes de la matrice de transition de la chaîne étaient identiques, ceci serait essentiellement équivalent à un modèle de mélange. La grande différence est que dans un *HMM*, le processus $\{C_t\}_{t\in\mathbb{N}^\star}$ n'est pas considéré *iid* par hypothèse. Cette petite différence aura cependant des répercussions très importantes sur les distributions conditionnelles et marginales de ces deux processus.

Dans les sections subséquentes, il sera parfois possible de démontrer certains résultats sous une définition plus générale des modèles de Markov à variables latentes dans laquelle la chaîne de Markov admet une dépendance entre C_t et $X_{1:t-1}$. Cette définition est présentée à la spécification généralisée (p.58).

Spécification 3 : HMM généralisé

Distribution initiale	$C_1 \sim p(c_1)$
Probabilité de transition	$C_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t-1} = c_{1:t-1}) \sim p(c_t \mid c_{t-1}, x_{1:t-1})$
Probabilité d'émission	$X_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t} = c_{1:t}) \sim p(x_t \mid x_{1:t-1}, c_t)$

Nouvelle notation associée

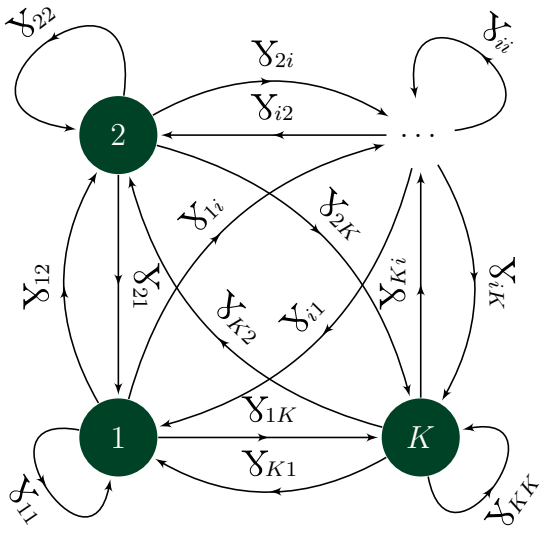
$$\vec{p}(x_t \mid x_{1:t-1}) \overset{\text{def}}{=} (p_1(x_t \mid x_{1:t-1}), \dots, p_K(x_t \mid x_{1:t-1})), \tag{3.1.3}$$

$$p_i(x_t \mid x_{1:t-1}) \overset{\text{def}}{=} \mathbb{P}[X_t = x_t \mid X_{1:t-1} = x_{1:t-1}, C_t = i], \tag{3.1.4}$$

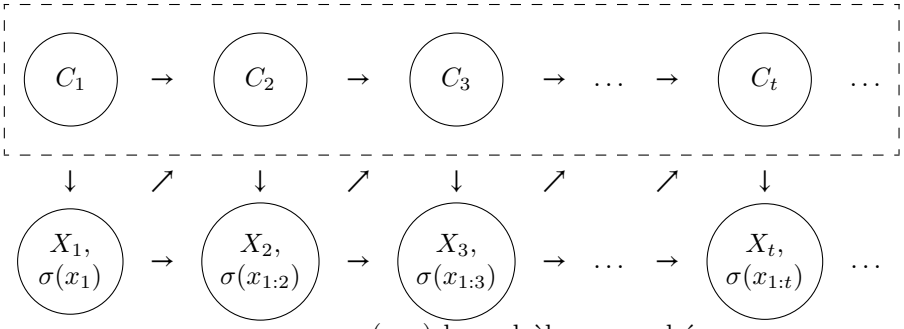
en plus de la matrice de transition qui dépend maintenant des valeurs passées du processus observé

$$\mathbf{\Gamma}_t = \Big(\mathbb{P}[C_{t+1} = j \mid C_t = i, x_{1:t}]\Big)_{i,j=1,\dots,K}. \tag{3.1.5}$$

Représentation graphique du processus latent



Représentation graphique du modèle complet



avec $\sigma(x_{1:t})$ la σ-algèbre engendrée par $x_{1:t}$.

3.2. Distributions marginales

3.2.1. Distributions marginales univariées

Les distributions marginales univariées du processus $\{X_t\}_{t\in\mathbb{N}^\star}$ sont obtenues en conditionnant sur les valeurs du processus $\{C_t\}_{t\in\mathbb{N}^\star}$. En reprenant la notation utilisée à la section 2.5, ces distributions sont données par

$$\mathbb{P}[X_t = x_t] = \sum_{i=1}^{K} \mathbb{P}[C_t = i]\mathbb{P}[X_t = x_t \mid C_t = i], \tag{3.2.1}$$

$$= \sum_{i=1}^{K} \mathfrak{u}_i(t) p_i(x_t), \tag{3.2.2}$$

$$= \vec{\mathfrak{u}}(t)\mathbf{P}(x_t)\mathbf{1}_K, \quad \forall\, t \in \mathbb{N}^\star, \tag{3.2.3}$$

avec $\vec{\mathfrak{u}}(t) = (\mathfrak{u}_1(t), \ldots, \mathfrak{u}_K(t))$ et $\mathbf{P}(x_t) = \text{diag}(p_1(x_t), \ldots, p_K(x_t))$. Ces distributions sont ainsi obtenues de la même manière qu'elles l'étaient pour les modèles de mélange à la section 1.3, à la différence qu'ici les probabilités de mélange sont spécifiquement les distributions marginales du processus $\{C_t\}_{t\in\mathbb{N}^\star}$. Il est donc attendu que les moments de X_t soient obtenus sensiblement par la même méthode que pour un modèle de mélange, à la différence qu'ici les probabilités de mélange sont dépendantes du temps t.

3.2.2. Distributions marginales de plus grand ordre

Afin de bien expliquer le calcul des distributions marginales de plus grand ordre, nous utiliserons l'approche proposée par Zucchini et Macdonald [2009] qui consiste à d'abord définir les *parents* de chaque variable impliquée. Ces *parents* correspondent en fait aux variables apparaissant plus tôt dans le processus stochastique et desquelles dépend la variable à l'étude. Ceci est équivalent à dire que les *parents* d'une variable sont les variables qui apparaissent dans le calcul de sa distribution marginale. Ainsi, une fois ces *parents* définis, la distribution jointe de variables aléatoires $V_{1:n}$ est donnée par

$$\mathbb{P}[V_{1:n}] = \prod_{i=1}^{n} \mathbb{P}[V_i \mid \text{pa}(V_i)], \tag{3.2.4}$$

où $\text{pa}(V_i)$ correspond à l'ensemble des *parents* de la variable aléatoire V_i.

Considérons par exemple les quatres variables aléatoires $\{X_t, X_{t+h}, C_t, C_{t+h}\}$. En suivant simplement les flèches sur le graphe orienté de la spécification standard (p.56), nous avons que

dans cet ensemble

$$\text{pa}(X_t) = \{C_t\}, \tag{3.2.5}$$

$$\text{pa}(X_{t+h}) = \{C_{t+h}\}, \tag{3.2.6}$$

$$\text{pa}(C_{t+h}) = \{C_t\} \tag{3.2.7}$$

et

$$\text{pa}(C_t) = \emptyset. \tag{3.2.8}$$

La distribution jointe de ces quatres variables aléatoires est alors

$$\mathbb{P}[X_t, X_{t+h}, C_t, C_{t+h}] = \mathbb{P}[C_t] \times \mathbb{P}[X_t \mid C_t] \times \mathbb{P}[C_{t+h} \mid C_t] \times \mathbb{P}[X_{t+h} \mid C_{t+h}]. \tag{3.2.9}$$

Les distributions marginales bivariées du processus $\{X_t\}_{t\in\mathbb{N}^\star}$ sont ensuite obtenues en sommant sur les valeurs possibles prisent par les parents de ces variables aéatoires.

Proposition 3.2.1. *Les distributions marginales bivariées—ou d'ordre deux—du processus observé d'un modèle de Markov à variables latentes non-homogène sont obtenues, en reprenant la notation utilisée à la section 2.5, en conditionnant sur les parents de chaque variable.*

$$\begin{aligned}
\mathbb{P}[X_t = x_t, X_{t+h} = x_{t+h}] &= \sum_{i,j=1}^{K} \mathbb{P}[X_t = x_t, X_{t+h} = x_{t+h}, C_t = i, C_{t+h} = j], \\
&= \sum_{i,j=1}^{K} \mathbb{P}[C_t = i]\mathbb{P}[X_t = x_t \mid C_t = i] \\
&\qquad \times \mathbb{P}[C_{t+h} = j \mid C_t = i]\mathbb{P}[X_{t+h} = x_{t+h} \mid C_{t+h} = j], \\
&= \sum_{i,j=1}^{K} \mathfrak{u}_i(t) p_i(x_t) \mathfrak{Y}_{ij}^{(h)}(t) p_j(x_{t+h}), \\
&= \vec{\mathfrak{u}}(t)\mathbf{P}(x_t)\mathbf{\Gamma}_t^{(h)}\mathbf{P}(x_{t+h})\mathbf{1}_K, \quad \forall\, t,h \in \mathbb{N}^\star.
\end{aligned} \tag{3.2.10}$$

Ce résultat est ensuite aisément étendu à des distributions marginales de plus grands ordres. Sans perte de généralité, prenons en exemple les distributions marginales trivariées. Celles-ci sont obtenues, $\forall\, t, h, p \in \mathbb{N}^\star$, en suivant la même approche que précédemment, soit

$$\mathbb{P}[X_t = x_t, X_{t+h} = x_{t+h}, X_{t+h+p} = x_{t+h+p}] = \vec{\mathfrak{u}}(t)\mathbf{P}(x_t)\mathbf{\Gamma}_t^{(h)}\mathbf{P}(x_{t+h})\mathbf{\Gamma}_{t+h}^{(p)}\mathbf{P}(x_{t+h+p})\mathbf{1}_K,$$

et ainsi de suite pour les ordres plus élevés.

Remarque 3.2.1. Nous reconnaissons dès lors la forme que prendra la vraisemblance d'un échantillon de T observations, soit $\mathcal{L}_T(\boldsymbol{\theta}) = \mathbb{P}[X_{1:T} = x_{1:T}]$. Celle-ci sera discutée aux sections 3.5 et 3.6.

3.3. Moments de $\{X_t\}_{t\in\mathbb{N}^\star}$

Afin d'obtenir le cas général, nous considérerons les moments d'une transformation $g(x)$ des variables aléatoires X_t. Ces transformations sont soumises à la condition que l'espérance conditionnelle aux états de la chaîne soit définie pour celles-ci. Cette généralisation, inspirée de Rydén, Teräsvirta et Åsbrink [1998](p. 222), nous permettra d'obtenir plus rapidement les moments du processus $\{X_t\}_{t\in\mathbb{N}^\star}$ et de tester la faculté de nos modèles à mouler les *faits stylisés* présentés à la section 1.1 (p.35).

Soit la notation suivante pour les moments conditionnels d'une transformation $g(x)$ de la variable aléatoire X_t, $\forall\, i \in \mathcal{C}$ et $\forall\, t \in \mathbb{N}^\star$:

$$\mu_{g|i} = \mathbb{E}[g(X_t) \mid C_t = i], \tag{3.3.1}$$

$$\sigma^2_{g|i} = \mathbb{V}ar[g(X_t) \mid C_t = i]. \tag{3.3.2}$$

Avec cette notation, nous avons que

$$\mathbb{E}[g(X_t)] = \sum_{i=1}^{K} \mathbb{E}[g(X_t) \mid C_t = i]\mathbb{P}[C_t = i], \tag{3.3.3}$$

$$= \sum_{i=1}^{K} \mathfrak{u}_i(t)\mathbb{E}[g(X_t) \mid C_t = i], \tag{3.3.4}$$

$$= \vec{\mathfrak{u}}(t)\mathbf{M}_g\,\mathbf{1}_K, \quad \forall t \in \mathbb{N}^\star, \tag{3.3.5}$$

avec $\mathbf{M}_g \overset{\text{def}}{=} \text{diag}\big(\boldsymbol{\mu}_g\big)$ et $\boldsymbol{\mu}_g \overset{\text{def}}{=} \big(\mu_{g|1}, \ldots, \mu_{g|K}\big)'$.

Remarque 3.3.1. Si la chaîne admet une distribution stationnaire $\vec{\pi}$ et que celle-ci est choisie comme distribution initiale, ce résultat est identique à celui obtenu pour un simple modèle de mélange (Section 1.3). En effet, par définition, nous avons que $\vec{\pi}\boldsymbol{\Gamma} = \vec{\pi}$, et donc que si $\vec{\mathfrak{u}}(1) \overset{\text{def}}{=} \vec{\pi}$, alors $\vec{\mathfrak{u}}(t) = \vec{\mathfrak{u}}(1) = \vec{\pi}$, $\forall t$.

Pour ce qui est des moments conjoints cependant, le fait que $\{C_t\}_{t\in\mathbb{N}^\star}$ soit une chaîne de Markov implique que $\forall\, t \in \mathbb{N}^\star$ et $\forall\, h \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}[g(X_t,X_{t+h})] &= \sum_{i,j=1}^{K} \mathbb{E}[g(X_t, X_{t+h}) \mid C_t = i, C_{t+h} = j]\mathbb{P}[C_t = i, C_{t+h} = j], \\
&= \sum_{i,j=1}^{K} \mathbb{E}[g(X_t, X_{t+h}) \mid C_t = i, C_{t+h} = j]\mathbb{P}[C_t = i]\mathbb{P}[C_{t+h} = j \mid C_t = i], \\
&= \sum_{i,j=1}^{K} \mathbb{E}[g(X_t, X_{t+h}) \mid C_t = i, C_{t+h} = j]\, \mathfrak{u}_i(t)\, \mathcal{Y}^{(h)}_{ij}(t).
\end{aligned} \tag{3.3.6}$$

À l'avenir, nous restreindrons notre analyse aux fonctions $g(x)$ pouvant être factorisées sous la forme $g(X_t, X_{t+h}) = g_1(X_t)g_2(X_{t+h})$. Cette restriction a pour effet que, puisque $\{X_t\}_{t\in\mathbb{N}^\star}$ est conditionnellement *iid* et que $C_t \notin \text{pa}(X_{t+h})$ (i.e. $\mathbb{P}[\,X_{t+h} \mid C_t, C_{t+h}\,] = \mathbb{P}[\,X_{t+h} \mid C_{t+h}\,]$), l'espérance précédente peut être réécrite sous la forme

$$\mathbb{E}[g(X_t, X_{t+h})] = \mathbb{E}[g_1(X_t)g_2(X_{t+h})], \tag{3.3.7}$$

$$= \sum_{i,j=1}^{K} \mathbb{E}[\,g_1(X_t)g_2(X_{t+h}) \mid C_t = i, C_{t+h} = j\,]\,\mathfrak{u}_i(t)\,\gimel_{ij}^{(h)}(t), \tag{3.3.8}$$

$$\overset{\text{ind}}{=} \sum_{i,j=1}^{K} \mathbb{E}[\,g_1(X_t) \mid C_t = i]\mathbb{E}[\,g_2(X_{t+h}) \mid C_{t+h} = j\,]\,\mathfrak{u}_i(t)\,\gimel_{ij}^{(h)}(t), \tag{3.3.9}$$

$$= \sum_{i,j=1}^{K} \mu_{g_1|i}\,\mu_{g_2|j}\,\mathfrak{u}_i(t)\,\gimel_{ij}^{(h)}(t), \tag{3.3.10}$$

$$= \vec{\mathfrak{u}}(t)\mathbf{M}_{g_1}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\mu}_{g_2}, \quad \forall t,h \in \mathbb{N}^\star, \tag{3.3.11}$$

avec $\mathbf{M}_{g_f} \overset{\text{def}}{=} \text{diag}\big(\boldsymbol{\mu}_{g_f}\big)$, $\boldsymbol{\mu}_{g_f} \overset{\text{def}}{=} \Big(\mu_{g_f|1}, \ldots, \mu_{g_f|K}\Big)'$ et $\mu_{g_f|i} \overset{\text{def}}{=} \mathbb{E}[g_f(X_t) \mid C_t = i], \quad \forall t \in \mathbb{N}^\star$.

Remarque 3.3.2. Cette forme générale pour les moments de X_t sera utile lorsque viendra le temps de calculer les moments marginaux et conjoints de la valeur absolue de notre variable. Nous pourrons ainsi tester la capacité de notre modèle à reproduire les *faits stylisés* (voir section 1.1). À cet effet, c'est plus spécifiquement la fonction d'autocovariance sous cette transformation qui nous intéressera. Celle-ci sera présentée à la section suivante.

Alors que la formule (3.3.6) est valide $\forall\, h \in \mathbb{N}$, la formule obtenue en (3.3.11) n'est pas toujours valide en $h = 0$. En général,

$$\mathbb{E}[g_1(X_t)g_2(X_t) \mid C_t = i] \neq \mathbb{E}[g_1(X_t) \mid C_t = i] \cdot \mathbb{E}[g_2(X_t) \mid C_t = i], \tag{3.3.12}$$

puisque habituellement $g_1(X_t)$ n'est pas indépendant de $g_2(X_t)$. En particulier, ceci sera problématique lorsque l'on cherche $\mathbb{E}[g(X_t)g(X_t) \mid C_t = i]$ pour calculer $\mathbb{V}ar[g(X_t)]$. Cependant, puisque nous savons que

$$\mathbb{V}ar[\,g(X_t) \mid C_t = i] = \mathbb{E}\Big[\,g(X_t)^2 \;\Big|\; C_t = i\Big] - (\mathbb{E}[\,g(X_t) \mid C_t = i])^2, \tag{3.3.13}$$

$$\Leftrightarrow \qquad \sigma_{g|i}^2 = \mu_{g^2|i} - \mu_{g|i}^2, \tag{3.3.14}$$

$$\Leftrightarrow \qquad \mu_{g^2|i} = \sigma_{g|i}^2 + \mu_{g|i}^2, \tag{3.3.15}$$

le second moment de $g(X_t)$ peut s'obtenir directement

$$\begin{aligned}
\mathbb{E}\Big[g(X_t)^2\Big] &= \mathbb{E}_{C_t}\Big[\mathbb{E}\Big[g(X_t)^2 \;\Big|\; C_t\Big]\Big] = \sum_{i=1}^{K}\mathbb{E}\Big[g(X_t)^2 \;\Big|\; C_t = i\Big]\mathbb{P}[C_t = i], \\
&= \sum_{i=1}^{K}\mu_{g^2|i}\,\mathfrak{u}_i(t), \\
&= \sum_{i=1}^{K}\Big(\sigma^2_{g|i} + \mu^2_{g|i}\Big)\mathfrak{u}_i(t), \\
&= \vec{\mathfrak{u}}(t)\Big(\mathbf{\Sigma}^2_g + \mathbf{M}^2_g\Big)\mathbf{1}_K, \qquad (3.3.16) \\
&= \vec{\mathfrak{u}}(t)\mathbf{M}_{g^2}\,\mathbf{1}_K, \qquad (3.3.17)
\end{aligned}$$

avec $\mathbf{M}_{g^2} = \text{diag}\Big(\boldsymbol{\mu}_{g^2}\Big)$, $\boldsymbol{\mu}_{g^2} = \Big(\mu_{g^2|1}, \ldots, \mu_{g^2|K}\Big)'$ et $\mu_{g^2|i} = \mathbb{E}\Big[g(X_t)^2 \;\Big|\; C_t = i\Big]$, $\forall\, t \in \mathbb{N}^\star$. Ainsi, nous pouvons finalement obtenir la formule pour la variance de $g(X_t)$ qui prend la forme

$$\begin{aligned}
\mathbb{V}ar[g(X_t)] &= \mathbb{E}\Big[g(X_t)^2\Big] - \big(\mathbb{E}[g(X_t)]\big)^2, \\
&= \vec{\mathfrak{u}}(t)\mathbf{M}_{g^2}\,\mathbf{1}_K - \big(\vec{\mathfrak{u}}(t)\mathbf{M}_g\,\mathbf{1}_K\big)^2, \qquad (3.3.18) \\
&= \vec{\mathfrak{u}}(t)\Big(\mathbf{\Sigma}^2_g + \mathbf{M}^2_g\Big)\mathbf{1}_K - \big(\vec{\mathfrak{u}}(t)\mathbf{M}_g\,\mathbf{1}_K\big)^2, \quad \forall t \in \mathbb{N}^\star. \qquad (3.3.19)
\end{aligned}$$

Remarque 3.3.3 (Formule de variance totale). Il est aussi possible d'obtenir la variance de $g(X_t)$ par la formule de variance totale. Le résultat obtenu est cependant bien moins élégant et ainsi nous le reportons dans l'appendice externe.

3.4. Autocovariance de $\{X_t\}_{t\in\mathbb{N}^\star}$ et $\{C_t\}_{t\in\mathbb{N}^\star}$

Grâce aux résultats (3.3.5) et (3.3.11), la fonction d'autocovariance de la variable aléatoire X_t s'obtient directement.

Proposition 3.4.1 (Autocovariance du processus $\{X_t\}_{t\in\mathbb{N}^\star}$)**.** *Soit $\{(X_t, C_t)\}_{t\in\mathbb{N}^\star}$, un modèle de Markov à variables latentes non-homogène tel que défini à la spécification standard* *(p.56)**. Pour tous entiers positifs quelconques t et h, la fonction d'autocovariance de $g_1(X_t)$ et $g_2(X_{t+h})$ est*

$$\begin{aligned}
\mathbb{C}ov[g_1(X_t), g_2(X_{t+h})] &= \mathbb{E}[g_1(X_t)g_2(X_{t+h})] - \mathbb{E}[g_1(X_t)] \times \mathbb{E}[g_2(X_{t+h})], \qquad (3.4.1) \\
&= \vec{\mathfrak{u}}(t)\mathbf{M}_{g_1}\mathbf{\Gamma}^{(h)}_t\boldsymbol{\mu}_{g_2} - \big(\vec{\mathfrak{u}}(t)\mathbf{M}_{g_1}\,\mathbf{1}_K\big) \times \big(\vec{\mathfrak{u}}(t+h)\mathbf{M}_{g_2}\,\mathbf{1}_K\big). \qquad (3.4.2)
\end{aligned}$$

En remplaçant les transformations g_1 et g_2 par la fonction identité $g(x) = x$, nous obtenons en particulier que

$$\mathbb{C}ov[X_t, X_{t+h}] = \vec{\mathfrak{u}}(t)\mathbf{M}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\mu} - (\vec{\mathfrak{u}}(t)\boldsymbol{\mu}) \times (\vec{\mathfrak{u}}(t+h)\boldsymbol{\mu}), \tag{3.4.3}$$

avec $\mathbf{M} \stackrel{\text{def}}{=} \text{diag}(\boldsymbol{\mu})$*, la matrice diagonale des moments conditionnels de* X_t*.*

Remarque 3.4.1. Il est intéressant de noter que la forme obtenue pour l'autocovariance de X_t est la même que celle associée à la variable C_t (présentée dans l'appendice externe). En effet, en remplaçant les entiers positifs comme étiquettes des régimes de la chaîne de Markov par les moments conditionnels de $\{X_t\}_{t\in\mathbb{N}^\star}$, nous obtenons que $\boldsymbol{\upsilon} \equiv \boldsymbol{\mu}$ et donc que $\boldsymbol{\Upsilon} \equiv \mathbf{M}$. Avec cette étiquettage, nous obtenons le même résultat en (3.4.3) que pour le processus $\{C_t\}_{t\in\mathbb{N}^\star}$, et nous avons que

$$\mathbb{C}ov[C_t, C_{t+h}] = \vec{\mathfrak{u}}(t)\boldsymbol{\Upsilon}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\upsilon} - (\vec{\mathfrak{u}}(t)\boldsymbol{\upsilon}) \times (\vec{\mathfrak{u}}(t+h)\boldsymbol{\upsilon})\Big|_{\boldsymbol{\upsilon}\equiv\boldsymbol{\mu}\,,\boldsymbol{\Upsilon}\equiv\mathbf{M}}\,, \tag{3.4.4}$$

$$= \vec{\mathfrak{u}}(t)\mathbf{M}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\mu} - (\vec{\mathfrak{u}}(t)\boldsymbol{\mu}) \times (\vec{\mathfrak{u}}(t+h)\boldsymbol{\mu}), \tag{3.4.5}$$

$$= \mathbb{C}ov[X_t, X_{t+h}], \quad \forall\, t,h \in \mathbb{N}^\star. \tag{3.4.6}$$

Le processus $\{X_t\}_{t\in\mathbb{N}^\star}$ a donc la même fonction d'autocovariance que la chaîne de Markov associée, sauf en $h = 0$.

Finalement, pour tous entiers positifs quelconques t et h, la fonction d'autocorrélation de $g_1(X_t)$ et $g_2(X_{t+h})$ est donnée par

$${}_{g_1,g_2}\rho_{X_t}^{(h)} = \mathbb{C}\text{orr}[g_1(X_t), g_2(X_{t+h})], \tag{3.4.7}$$

$$= \frac{\mathbb{C}ov[g_1(X_t), g_2(X_{t+h})]}{\sqrt{\mathbb{V}ar[g_1(X_t)]}\sqrt{\mathbb{V}ar[g_2(X_{t+h})]}}. \tag{3.4.8}$$

En particulier, nous avons que la fonction d'autocorrélation de X_t est

$$\rho_{X_t}^{(h)} = \mathbb{C}\text{orr}[X_t, X_{t+h}], \tag{3.4.9}$$

$$= \frac{\mathbb{C}ov[X_t, X_{t+h}]}{\sqrt{\mathbb{V}ar[X_t]}\sqrt{\mathbb{V}ar[X_{t+h}]}}. \tag{3.4.10}$$

3.5. Log-vraisemblance

3.5.1. Calcul sur toutes les trajectoires possibles

Il pourrait être tentant de calculer directement la vraisemblance d'un échantillon $x_{1:T}$ en conditionnant sur toutes les valeurs possibles que puissent prendre la chaîne de Markov

latente. Cette vraisemblance serait alors donnée par

$$\begin{aligned}\mathcal{L}_T(\boldsymbol{\theta}) &= \sum_{c_{1:T}} \mathbb{P}[x_{1:T}, c_{1:T}], \\ &= \sum_{c_{1:T}} \mathbb{P}[x_{1:T} \mid c_{1:T}]\mathbb{P}[c_{1:T}], \\ &= \sum_{c_{1:T}} \left[\left(\prod_{t=1}^{T} \mathbb{P}[x_t \mid c_t] \right) \left(\mathbb{P}[c_1] \prod_{t=2}^{T} \mathbb{P}[c_t \mid c_{t-1}] \right) \right]\end{aligned}$$

En considérant un *HMM* avec espace d'états $\mathcal{C}$ de cardinalité K sur une chaîne de longueur T, il existe K^T combinaisons possibles de ces états, correspondant chacune à un produit de $2T$ éléments. Ce calcul demande donc $\mathrm{O}\left(TK^T\right)$ opérations.

Afin d'illustrer l'énormité de cette valeur, penchons-nous sur le calcul du nombre d'atomes dans l'univers visible. Il est estimé que notre étoile, le Soleil, ait une masse de $\approx 2^{33}g$, correspondant à $\approx 10^{57}$ atomes. En utilisant cette valeur comme moyenne pour la taille des étoiles ailleurs dans l'univers et en multipliant par le nombre estimé d'étoiles dans notre galaxie (100 milliards = 10^{11}), nous obtenons que notre galaxie compterait $\approx 10^{68}$ atomes. Puisque notre galaxie est considérée de taille moyenne, nous pouvons donc la prendre comme galaxie-type et multiplier sa taille par le nombre estimé de galaxie dans l'univers visible (1000 milliards = 10^{12}). Tous ces calculs nous emmènent à estimer qu'il y aurait, dans l'univers visible, $\approx 10^{68} \times 10^{12} = 10^{80} \approx 2^{265.7542}$ atomes.

Cette valeur peut sembler élevée—et elle l'est—, mais elle correspond en fait au nombre de combinaisons possibles pour un modèle très simple avec $K = \mathrm{Card}(\mathcal{C}) = 2$, et seulement $T = 266$ observations. Une telle chaîne serait considérée très courte, voir minuscule, en finance. C'est en fait à peine supérieur au nombre prévu de jours ouvrables pour la bourse de Toronto en 2020, c'est-à-dire 252 [3]. Il est donc clair que cette voie ne soit pas computationnellement admissible comme solution, même pour le modèle le plus simple à 2 états.

S'il n'est pas possible de calculer la vraisemblance en sommant sur toutes les valeurs possibles de la chaîne, il faudra pouvoir obtenir de l'information sur le processus latent $\{C_t\}_{t\in\mathbb{N}^\star}$ et réduire la cardinalité de l'espace exploré. C'est précisément cette recherche d'information qui est l'objet du problème de filtrage, un défi classique des modèles stochastiques.

3.5.2. Problème de filtrage

Afin d'expliquer le problème de filtrage, il convient de présenter rapidement le concept de filtration. À la jonction de la théorie des ensembles et des processus stochastiques et

3. 366 jours par année bissextile - 104 jours de fin de semaine - 10 journées fériées = 252 jours ouvrables. (Wikipedia contributors [2019] et TSX [2020])

introduites dans les travaux de Joseph Doob au milieu du 20e siècle, les filtrations permettent de formaliser la notion d'information observée du point 1 jusqu'à un certain point t dans le temps inclusivement.

Définition 3.5.1 (Filtrations)**.** *En théorie des processus stochastiques, une filtration sur un ensemble est une suite de parties croissantes pour l'inclusion. Cette famille indexée, notée* $\{\boldsymbol{\mathcal{F}}_t\}_{t\in\mathbb{N}^\star}$*, est engendrée par le processus stochastique* $\{X_t\}_{t\in\mathbb{N}^\star}$*. Chacuns des ensembles* $\boldsymbol{\mathcal{F}}_m$ *de cette suite contient l'information observée jusqu'au point* m *inclusivement. De plus, ces ensembles respectent la condition que*

$$\boldsymbol{\mathcal{F}}_k \subseteq \boldsymbol{\mathcal{F}}_l, \quad \forall\, k \leq l \in \mathbb{N}^\star, \tag{3.5.1}$$

avec la précision que toutes les suites de $\{\boldsymbol{\mathcal{F}}_t\}_{t>n}$ *partagent les mêmes* n *premiers termes.*

Cette définition n'est pas une définition formelle car celle-ci aurait usage de concepts qui ne sont pas l'objet de ce mémoire et qui n'aurait pas de valeur ajoutée en terme de clareté du texte. Pour une définition formelle des filtrations, les travaux de Joseph Doob—en particulier son livre Doob [1953]—ou le papier de Coculescu et Nikeghbali [2007] traitent plus en profondeur des diverses caractéristiques, conséquences et utilisations de celles-ci.

Maintenant que le concept de filtration est défini, précisons ce qui est entendu par *le problème de filtrage*. Supposons d'abord que notre modèle soit complètement spécifié et que nous ayons mesuré l'information observable jusqu'au temps t inclusivement. Dans ce cas, le problème de filtrage revient à poser la question :

« Qu'est-il possible de dire sur le processus latent $\{C_t\}_{t\in\mathbb{N}^\star}$, conditionnellement à la filtration $\{\boldsymbol{\mathcal{F}}_t\}_{t\in\mathbb{N}^\star}$? »,

ou de manière équivalente :

« Comment est-il possible de calculer la probabilité conditionnelle $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$? ».

Dans le cadre d'un modèle standard, nous avons que $\boldsymbol{\mathcal{F}}_t$ correspond à l'information engendrée par la suite $x_{1:t}$. Ceci signifie qu'en terme mathématique, le problème de filtrage se résume à obtenir $\mathbb{P}[C_t \mid x_{1:t}], \forall\, t \in \mathbb{N}^\star$. La prochaine étape est donc de proposer une méthode afin d'inférer sur les états cachés de la chaîne de Markov, et ce, en conditionnant sur l'information contenue dans la suite $\{\boldsymbol{\mathcal{F}}_t\}_{t\in\mathbb{N}^\star}$.

3.6. Inférence sur les états cachés de la chaîne

Dans cette section, nous nous intéressons aux méthodes permettant de faire de l'inférence sur l'état caché de la chaîne de Markov au temps t, en conditionnant sur l'information observée jusqu'au temps t, soit $\boldsymbol{\mathcal{F}}_t$. Une première méthode, dénommée l'**algorithme avant** par

Lystig et Hughes [2002] et présentée initialement dans Baum, Petrie, Soules et Weiss [1970], permet en théorie de calculer efficacement la vraisemblance en calculant récursivement les probabilités $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$, dites **probabilités filtrées**.

Bémol important cependant, cet algorithme souffre du problème de soupassement arithmétique (*underflow*). Nous présenterons donc directement une technique connue sous le nom de **filtre d'hamilton**—ou **algorithme de filtrage avant**—permettant d'éviter ce problème de soupassement. Le nom de cet algorithme est lié aux travaux de Hamilton [1989 ; 1994 ; 2008], même s'il n'était pas le premier à proposer une telle méthode. L'algorithme avant, ainsi que la preuve de celui-ci, se retrouve malgré tout dans l'appendice externe.

Toujours en conditionnant sur $\boldsymbol{\mathcal{F}}_t$, nous présenterons ensuite une technique permettant de faire des prévisions sur les états futurs de la chaîne, soit en $t + h$.

Remarque 3.6.1 (Inférence conditionnelle à $\boldsymbol{\mathcal{F}}_T$). Il est aussi possible de faire de l'inférence sur les états cachés de la chaîne en conditionnant sur toute l'information disponible jusqu'à la fin de la fenêtre d'observation, soit $\boldsymbol{\mathcal{F}}_T$. Le **filtre de Kim** permet dans ce contexte d'obtenir les probabilités $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_T]$ dites **probabilités lissées**, et l'**algorithme de Viterbi** permet d'obtenir la séquence d'états cachés la plus probable a posteriori. Ces quantités n'étant cependant pas centrales à l'objet de ce mémoire, elles sont reportées dans l'appendice externe.

3.6.1. Inférence conditionnelle à $\boldsymbol{\mathcal{F}}_t$

3.6.1.1. Filtre d'Hamilton

En plus d'être une solution au problème de soupassement arithmétique et de dépassement d'entier, cet algorithme permet d'obtenir les trois termes suivants $\forall\, t \in \{1, \ldots, T\}$:

$$\begin{aligned} \text{Densité conditionnelle} \quad & \mathbb{P}[X_t \mid X_{1:t-1}], \\ \text{Probabilité filtrée} \quad & \mathbb{P}[C_t \mid X_{1:t}], \\ \text{Probabilité prédictive} \quad & \mathbb{P}[C_{t+1} \mid X_{1:t}]. \end{aligned}$$

Ces probabilités seront largement utilisées dans les sections subséquentes, en particulier lorsqu'il sera question de prédiction à la section 3.6.1.2. Ces probabilités sont aussi cruciales à l'exécution de l'algorithme de lissage présentés dans l'appendice externe.

Tout comme l'algorithme avant, il est intéressant de noter que le filtre d'Hamilton permet d'obtenir la log-vraisemblance pour un *HMM* sous la spécification généralisée (p.58), soit un modèle de Markov à variables latentes dont la chaîne de Markov admet une dépendance entre C_t et $X_{1:t-1}$.

Remarque 3.6.2. Le filtre d'Hamilton permet en fait une généralisation encore plus forte de notre modèle en admettant une dépendance entre C_t et $C_{t-q:t-1}$, soit un modèle de Markov

à variables latentes ayant une chaîne de Markov d'ordre $q \in \mathbb{N}^\star$. Cette relaxation alourdie énormément le modèle cependant, nécessitant $\mathrm{O}(TK^{q+1})$ opérations et n'est donc pas pratiquable, sauf pour de petits échantillons avec une chaîne d'ordre q relativement petit. Voir Augustyniak [2018](p. 85-86) pour plus de détails ainsi que la méthode du filtre d'Hamilton généralisé.

Puisque le filtre d'Hamilton est très important, et afin de bien présenter chaque étape du filtre et de son implémentation, nous allons d'abord présenter l'approche probabiliste, puis une version algorithmique de ce filtre plus facile à implanter dans un logiciel de programmation. Malgré le fait que les deux approches soient équivalentes et qu'elles comportent toutes deux trois étapes, ces étapes diffèrent. Une version probabiliste est présentée à la figure 3.1, et une version algorithmique est présentée dans l'algorithme 1 : Filtre d'Hamilton (p.68).

Algorithme 1 : Filtre d'Hamilton

Soit les notations suivantes, et leurs interpretations probabilistes :

$$\vec{a}_t \overset{\text{def}}{=} (\mathbb{P}[X_t, C_t = 1 \mid X_{1:t-1}], \ldots, \mathbb{P}[X_t, C_t = K \mid X_{1:t-1}]), \tag{3.6.1}$$

$$a_t \overset{\text{def}}{=} \mathbb{P}[X_t \mid X_{1:t-1}], \tag{3.6.2}$$

$$\vec{\omega}_t \overset{\text{def}}{=} (\mathbb{P}[C_t = 1 \mid X_{1:t}], \ldots, \mathbb{P}[C_t = K \mid X_{1:t}]). \tag{3.6.3}$$

Étape 1

$$\vec{a}_1 = \vec{\mathfrak{u}}(1) \odot \vec{p}(x_1),$$

$$a_1 = \vec{a}_1 \mathbf{1}_K,$$

$$\vec{\omega}_1 = {}^{\vec{a}_1}/_{a_1}.$$

Étape 2 : $t = 2, \ldots$

$$\vec{a}_t = (\vec{\omega}_{t-1} \mathbf{\Gamma}_{t-1}) \odot \vec{p}(x_t \mid x_{1:t-1}),$$

$$a_t = \vec{a}_t \mathbf{1}_K,$$

$$\vec{\omega}_t = {}^{\vec{a}_t}/_{a_t}.$$

Étape 3

$$\ell_T(\boldsymbol{\theta}) \quad = \quad \sum_{t=1}^{T} \log(a_t)$$

Coût computationnel	Référence
$\mathrm{O}(TK^2)$	Augustyniak [2018](p. 82-86)

Remarque

Cet algorithme est valide même pour un modèle de Markov à variables latentes non-homogène tel que défini à la spécification généralisée (p.58).

Étape 1

$$\mathbb{P}\Big[\, X_1 \,\Big] = \sum_{i=1}^{K} \mathbb{P}[X_1, C_1 = i] = \sum_{i=1}^{K} \mathbb{P}[X_1 \mid C_1 = i]\mathbb{P}[C_1 = i]$$

$$\mathbb{P}\Big[\, C_1 \;\Big|\; X_1 \,\Big] = \frac{\mathbb{P}[X_1, C_1]}{\mathbb{P}[X_1]} = \frac{\mathbb{P}[X_1, C_1]}{\sum\limits_{i=1}^{K} \mathbb{P}[X_1, C_1 = i]}$$

$$\mathbb{P}\Big[\, C_2 \;\Big|\; X_1 \,\Big] = \sum_{i=1}^{K} \mathbb{P}[C_2, C_1 = i \mid X_1] = \sum_{i=1}^{K} \mathbb{P}[C_2 \mid C_1 = i, X_1]\mathbb{P}[C_1 = i \mid X_1]$$

Étape 2 - $t = 2, \ldots, T$

$$\begin{aligned}\mathbb{P}\Big[\, X_t \;\Big|\; X_{1:t-1} \,\Big] &= \sum_{i=1}^{K} \mathbb{P}[X_t, C_t = i \mid X_{1:t-1}], \\ &= \sum_{i=1}^{K} \mathbb{P}[X_t \mid X_{1:t-1}, C_t = i]\mathbb{P}[C_t = i \mid X_{1:t-1}]\end{aligned}$$

$$\mathbb{P}\Big[\, C_t \;\Big|\; X_{1:t} \,\Big] = \frac{\mathbb{P}[X_t, C_t \mid X_{1:t-1}]}{\mathbb{P}[X_t \mid X_{1:t-1}]} = \frac{\mathbb{P}[X_t \mid X_{1:t-1}, C_t]\mathbb{P}[C_t \mid X_{1:t-1}]}{\mathbb{P}[X_t \mid X_{1:t-1}]}$$

$$\begin{aligned}\mathbb{P}\Big[\, C_{t+1} \;\Big|\; X_{1:t} \,\Big] &= \sum_{i=1}^{K} \mathbb{P}[C_{t+1}, C_t = i \mid X_{1:t}], \\ &= \sum_{i=1}^{K} \mathbb{P}[C_{t+1} \mid C_t = i, X_{1:t}]\mathbb{P}[C_t = i \mid X_{1:t}]\end{aligned}$$

Étape 3

$$\ell_T(\boldsymbol{\theta}) \;=\; \log(\mathbb{P}[X_{1:T}]) \;=\; \log(\mathbb{P}[X_1]) \;+\; \sum_{t=2}^{T} \log(\mathbb{P}[X_t \mid X_{1:t-1}]).$$

Figure 3.1 – Approche probabiliste du filtre d'Hamilton

Remarque 3.6.3 (Échantillons incomplets). Une des caractéristiques attrayantes de cet algorithme est qu'il s'adapte très facilement lorsque l'échantillon est incomplet. En effet, il suffit alors de remplacer la matrice diagonale des distributions conditionnelles $\mathbf{P}(x_t)$ par la matrice identité. Ceci est équivalent à omettre cette matrice dans le calcul et à remplacer $\mathbf{\Gamma}_t$ par $\mathbf{\Gamma}_t^{(1+m)}$, où m correspond au nombre de données manquantes consécutives. Bien entendu, il ne faut pas que la matrice de transition dépende de ces valeurs manquantes, ce qui signifie

que nous devons exclure la spécification généralisée (p.58) et nous en tenir à la spécification standard (p.56) lorsque des données sont manquantes.

En supposant que les observations indicées $\{3, 4\}$ soient manquantes, la vraisemblance devient

$$\mathcal{L}_T^{-(3,4)}(\boldsymbol{\theta}) = \vec{\mathfrak{u}}(1)\mathbf{P}(x_1)\boldsymbol{\Gamma}_1\mathbf{P}(x_2)\boldsymbol{\Gamma}_2^{(3)}\mathbf{P}(x_5)\boldsymbol{\Gamma}_5 \dots \boldsymbol{\Gamma}_{T-1}\mathbf{P}(x_T)\mathbf{1}_K, \tag{3.6.4}$$

le reste de l'algorithme restant inchangé.

3.6.1.2. Inférence en $t + h$

Faire des prévisions sur les états cachés de la chaîne de Markov en $(t + h)$ conditionnellement à $\boldsymbol{\mathcal{F}}_t$ nécessite, sous notre modèle actuelle, de trouver une expression pour calculer $\mathbb{P}[C_{t+h} \mid \boldsymbol{\mathcal{F}}_t]$.

Tout d'abord, il est toujours possible de prédire l'état de la chaîne de Markov un pas en avant (i.e. en $h = 1$). La valeur de $\mathbb{P}[C_{t+1} \mid \boldsymbol{\mathcal{F}}_t]$ est en fait obtenue directement par le biais du filtre d'Hamilton (3.6.1.1). C'est d'ailleurs aussi grâce à ce filtre que nous obtenons les termes nécessaires à la prévision de l'état de la chaîne en $(t + h)$.

Pour ce qui est des prévisions plus d'un pas en avant (i.e. en $h \geq 2$), l'homogénéité de la matrice de transition est alors une propriété très intéressante, voir *sine qua non*. En effet, lorsque la matrice de transition de la chaîne de Markov est homogène—et donc unique—, il est possible d'utiliser l'information observée jusqu'à t pour inférer analytiquement sur tous les états subséquents de la chaîne de Markov. Ceci est une conséquence directe d'une propriété des matrices de transition en m pas (voir l'appendice externe).

Si toutefois la matrice de transition n'est pas homogène, une inférence analytique n'est pas forcément possible. Lorsque par exemple les entrées de la matrice de transition en h pas dépendent de valeurs n'ayant pas encore été observées, il n'est pas possible d'obtenir analytiquement la valeur de la matrice $\boldsymbol{\Gamma}_t^{(h)}$. Il faudra alors recourir à la simulation afin d'inférer sur ces valeurs au delà de $h = 1$. C'est le cas pour les *HMMs* sous la spécification généralisée (p.58) puisque la matrice de transition $\boldsymbol{\Gamma}_t^{(h)}$ dépend des valeurs du processus observé jusqu'à $(t + h - 1)$.

Puisque plusieurs architectures non-homogènes pour la matrice de transition seront présentées au chapitre 5, nous y réservons la discussion concernant les répercussions de celles-ci sur la prévision. Nous nous concentrerons pour le moment sur les matrices de transition homogènes, et celles qui ne dépendent que de valeurs incluses dans $\boldsymbol{\mathcal{F}}_t$, ou à tout le moins connues en t.

Proposition 3.6.1 (Prévision dans le processus $\{C_t\}_{t\in\mathbb{N}^\star}$)**.** *Soit $\{(X_t, C_t)\}_{t\in\mathbb{N}^\star}$, un modèle de Markov à variables latentes homogène tel que défini à la spécification standard (p.56). Nous avons alors, $\forall\, h \in \mathbb{N}^\star$, les probabilités*

$$\vec{\omega}_{t+h\,|\,t}(j) \stackrel{\text{def}}{=} \mathbb{P}[C_{t+h} = j \mid \boldsymbol{\mathcal{F}}_t] = \sum_{i=1}^{K} \mathbb{P}[C_t = i, C_{t+h} = j \mid \boldsymbol{\mathcal{F}}_t], \tag{3.6.5}$$

$$= \sum_{i=1}^{K} \mathbb{P}[C_t = i \mid \boldsymbol{\mathcal{F}}_t]\mathbb{P}[C_{t+h} = j \mid C_t = i, \boldsymbol{\mathcal{F}}_t], \tag{3.6.6}$$

$$= \sum_{i=1}^{K} \mathbb{P}[C_t = i \mid \boldsymbol{\mathcal{F}}_t]\mathbb{P}[C_{t+h} = j \mid C_t = i], \quad \textit{par homogénéité,} \tag{3.6.7}$$

$$= \sum_{i=1}^{K} \mathbb{P}[C_t = i \mid \boldsymbol{\mathcal{F}}_t]\gimel_{ij}^{(h)}, j = 1, \ldots, K. \tag{3.6.8}$$

avec $\boldsymbol{\mathcal{F}}_t \stackrel{\text{def}}{=} \sigma(x_{1:t})$.

En utilisant à nouveau la notation présentée dans le cadre du filtre d'Hamilton, les probabilités conditionnelles peuvent être exprimées sous forme matricielle :

$$\vec{\omega}_{t+h\,|\,t} \stackrel{\text{def}}{=} (\mathbb{P}[C_{t+h} = 1 \mid \boldsymbol{\mathcal{F}}_t], \ldots, \mathbb{P}[C_{t+h} = K \mid \boldsymbol{\mathcal{F}}_t]) = \vec{\omega}_t \boldsymbol{\Gamma}^h, \quad \forall\, h \in \mathbb{N}^\star. \tag{3.6.9}$$

Faire une prévision revient alors à prendre le maximum sur le vecteur de probabilités résultant.

3.7. Prévision sur le processus observé

À la section 3.6.1.2, nous présentions les méthodes permettant d'inférer les états cachés de la chaîne en $t+h$ conditionnellement à $\boldsymbol{\mathcal{F}}_t$. Si nous voulons maintenant prédire les valeurs de X_{t+h} pour $h \in \mathbb{N}^\star$, nous chercherons à calculer les probabilités conditionnelles

$$\mathbb{P}[x_{t+h} \mid x_{1:t}]. \tag{3.7.1}$$

Le calcul de ces probabilités en $h = 1$ est très simple car les densités prédictives un pas en avant sont obtenues directement à partir du filtre d'Hamilton.

Pour $h = 1$

$$\begin{aligned}\mathbb{P}[x_{t+1} \mid x_{1:t}] &= \sum_{j=1}^{K} \mathbb{P}[x_{t+1}, C_{t+1} = j \mid x_{1:t}], \\ &= \sum_{j=1}^{K} \mathbb{P}[C_{t+1} = j \mid x_{1:t}]\mathbb{P}[x_{t+1} \mid x_{1:t}, C_{t+1} = j], \qquad (3.7.2) \\ &= \sum_{j=1}^{K} \vec{\omega}_{t+1\,|\,t}(j) p_j(x_{t+1} \mid x_{1:t}). \qquad (3.7.3)\end{aligned}$$

À nouveau, lorsque la matrice de transition n'est pas homogène, il n'est habituellement pas possible d'obtenir analytiquement les densités prédictive pour $h \geq 2$. Il serait alors nécessaire de procéder à des simulations afin d'inférer les probabilités $\mathbb{P}[C_{t+h} \mid C_t, x_{1:t}]$, car nous ne connaissons pas les valeurs prisent par $X_{t+1:t+h}$—ni les valeurs prisent par les variables explicatives potentiellement utilisées dans le calcul de $\mathbf{\Gamma}_t^{(h)}$.

En contrepartie, sous la spécification standard (p.56), ces prévisions sont possibles et la méthode présentée ci-bas permet d'obtenir celles-ci analytiquement.

Pour $h \geq 2$

$$\begin{aligned}\mathbb{P}[x_{t+h} \mid x_{1:t}] &= \sum_{j=1}^{K} \mathbb{P}[x_{t+h}, C_{t+h} = j \mid x_{1:t}], \\ &= \sum_{j=1}^{K} \mathbb{P}[C_{t+h} = j \mid x_{1:t}]\mathbb{P}[x_{t+h} \mid x_{1:t}, C_{t+h} = j], \\ &= \sum_{j=1}^{K} \mathbb{P}[C_{t+h} = j \mid x_{1:t}]\mathbb{P}[x_{t+h} \mid C_{t+h} = j], \\ &= \sum_{j=1}^{K} \vec{\omega}_{t+h\,|\,t}(j) p_j(x_{t+h}). \qquad (3.7.4)\end{aligned}$$

En utilisant la notation vectorielle

$$\vec{\omega}_{t+h\,|\,t} = (\mathbb{P}[C_{t+h} = 1 \mid x_{1:t}], \ldots, \mathbb{P}[C_{t+h} = K \mid x_{1:t}]), \quad h \in \mathbb{N}^\star \qquad (3.7.5)$$

proposée à la section 3.6.1.2, ainsi que le résultat (3.6.9) de la proposition 3.6.1, il est possible de reformuler les relations (3.7.3) et (3.7.4) sous forme matricielle. En effet, nous avons alors

$$\begin{aligned}\mathbb{P}[x_{t+h} \mid x_{1:t}] &= \vec{\omega}_{t+h\,|\,t} \mathbf{P}(x_{t+h}) \mathbf{1}_K, \qquad (3.7.6) \\ &= \vec{\omega}_t \mathbf{\Gamma}^h \mathbf{P}(x_{t+h}) \mathbf{1}_K, \quad \forall\, h \in \mathbb{N}^\star, \qquad (3.7.7)\end{aligned}$$

où $\vec{\omega}_t$ contient les probabilités $\mathbb{P}[C_t = i \mid x_{1:t}]$, $i = 1, \ldots, K$, obtenues avec le filtre d'Hamilton—voir l'algorithme 1 : Filtre d'Hamilton (p.68).

Maintenant que nous avons défini les propriétés et discuté des conséquences des modèles de Markov à variables latentes, nous pouvons nous intéresser à différentes architectures possibles du processus de changement de régime, en commençant par la reformulation hiérarchique des *HMMs* présentée au chapitre suivant.

Chapitre 4

Reformulation hiérarchique - HHMM

Les modèles de Markov à variables latentes hiérarchiques —abbréviés *HHMMs* —, sont initialiement présentés dans les travaux de Fine, Singer et Tishby [1998] afin d'améliorer la modélisation du langage naturel. Leur objectif était alors d'incorporer à cette modélisation l'information contenue dans les structures plus simples (i.e quelques mots), mais sans négliger l'influence de structures plus complexes (i.e. des phrases, paragraphes, etc). Dans les recherches subséquentes, cette reformulation fût mise en valeur pour sa faculté à augmenter la granularité des modèles, tout en restant fidèle à l'idée derrière les *HMMs*.

Essentiellement, les *HHMMs* sont une reformulation multi-couche structurée des modèles de Markov à variables latentes. À chaque état d'un *HHMM* est en général associé un processus stochatisque à part entière. Certains de ces états n'ont d'influence que sur le reste du processus latent, et les états de la dernière couche émettent des observations, à l'instar des *HMMs* classiques. Sous cette reformulation, il est possible de voir chaque régime de l'espace d'états d'une couche comme étant une chaîne de Markov à son tour. Ainsi, les *HMMs* peuvent être vus comme un cas particulier des *HHMMs* dans lesquels chaque état de la chaîne est associé à une chaîne de Markov n'admettant qu'un état. C'est donc de dire que tout *HMM* est équivalent à un *HHMM* à une couche.

Remarque 4.0.1. Il est important de souligner à cet égard que tout *HHMM* peut aussi être représenté comme un *HMM* standard. Cette caractéristique sera évidente lorsque nous présenterons l'architecture de la matrice de transition pour les deux structures étudiées dans ce chapitre.

Pour ce qui est de choisir une distribution d'émission sous ces modèles, la procédure de sélection est, comme son nom l'indique, hiérarchisée—celle-ci produisant plutôt une série d'états qu'un état unique. Dans un premier temps, la couche supérieure émet un état, qui est ensuite envoyé à la couche suivante. Dans cette dernière, un second état est émis, ajouté à la série, et à nouveau cette série est envoyée à la couche suivante, et ainsi de suite jusqu'à

la dernière couche. Cette dernière couche finalise ensuite la construction de la série en y ajoutant un régime et, en général, émet un régime *producteur* (*production state*[1][2]) qui pointe finalement vers une distribution d'émission.

Remarque 4.0.2 (Modèles de Markov à variables latentes factoriels). Certaines versions des *HHMMs* soumettent la série d'états finale à une transformation, plutôt que de lui faire directement correspondre une distribution. C'est le cas par exemple des modèles de Markov à variables latentes factoriels (*Factorial Hidden Markov Model* ou *FHMM*), dans lesquels les états émis par les chaînes de Markov sont multipliés ensemble plutôt que concaténés. En effet, sous cette spécification, les états émis par les couches successives sont multipliés ensemble pour produire une valeur qui peut ensuite servir à mettre à l'échelle la variance du processus, comme par exemple dans Augustyniak, Bauwens et Dufays [2019].

Sous sa structure générale, c'est la granularité du modèle qui est la principale caractéristique recherchée. Au delà de cette caratéristique, ces modèles permettent aussi de visualiser le modèle sous un nouvel angle, en remplaçant les probabilités de transition par des fonctions déterministes de nouveaux paramètres pour les transitions. En revanche, depuis plusieurs années, l'accent de la recherche a été mis sur le développement de modèles étant plus parcimonieux (i.e. ayant un nombre restreint de paramètres). C'est en effet une quête double de parcimonie, mais aussi d'interpretabilité des régimes, qui motivent les scientifiques à étudier des modèles ayant entre deux et quatre régimes *producteurs*, et nécessitant moins de paramètres que les *HMMs* classiques.

Afin d'étudier à la fois la faculté des *HHMMs* à augmenter la granularité des modèles et leur faculté à augmenter leur parcimonie—tout en ne nous étendant pas trop sur le sujet—nous restreindrons notre étude à deux de ces structures. Dans un premier temps, nous étudierons une structure plus générale à la section 4.1, puis une structure simplifiée plus parcimonieuse sera présentée à la section 4.2.

Dans ces sections, nous chercherons à présenter les diverses spécifications de ces modèles aussi généralement que possible, mais nous présenterons seulement les résultats pour les modèles ayant deux couches cachées admettant chacune deux régimes possibles, soit un total de 4 régimes *producteurs*. À cet effet, nous pouvons dès maintenant définir la notation permettant d'identifier les diverses variables aléatoires impliquées dans ce type de modèle. Puisque notre modèle n'a que deux couches, nous choisissons de donner aux régimes de la couche supérieure le titre de régimes *parents* et à ceux de la couche subalterne le titre de régimes *enfants*. La série de régimes *parents* sera représentée par $\{P_t\}_{t\in\mathbb{N}^\star}$ et celle de régimes *enfants* par $\{E_t\}_{t\in\mathbb{N}^\star}$, avec la combinaison de ces deux séries permettant de constituer la série

1. Charlot et Marimoutou [2011].
2. Augustyniak, Bauwens et Dufays [2019].

de régimes *producteurs* $\{C_t\}_{t\in\mathbb{N}^\star} \equiv \{(P_t, E_t)\}_{t\in\mathbb{N}^\star}$. Nous réservons pour les sections suivantes la discussion concernant les valeurs que peuvent prendre ces variables.

4.1. Structure générale

La structure générale des modèles de Markov à variables latentes hiérarchiques, inspirée des travaux de Charlot et Marimoutou [2011], est présentée à la spécification hiérarchique générale (p.78). Les représentations graphiques incluses dans cette spécification sont associées à un *HHMM* ayant deux couches, et où les chaînes de la couche subalterne ne partagent pas forcément le même nombre de régimes. Cette structure est aisément étendue à des modèles ayant plus de couches, la figure présentée à la portion *représentation graphique du processus latent* illustrant simplement la relation entre les régimes d'une couche supérieure (i.e. les régimes *parents*) et de la couche subalterne suivante (i.e. les régimes *enfants* associés).

Dans le cas général d'un modèle avec N régimes *parents* ayant respectivement k_i régimes *enfants*, les régimes *parents* peuvent prendre les valeurs $\{p_1, \ldots, p_N\}$ et les régimes *enfants* du i^{e} parent les valeurs $\left\{e_1^i, \ldots, e_{k_i}^i\right\}$. Les régimes *producteurs* sont alors issus de l'espace d'états $\mathcal{C} = \left\{(p_1, e_1^1), \ldots, \left(p_1, e_{k_1}^1\right), \ldots, \left(p_N, e_1^N\right), \ldots, \left(p_N, e_{k_N}^N\right)\right\}$. Ainsi, à un pas de temps t quelconque, le régime *parent*

$$P_t \in \overset{\triangle}{\mathcal{C}} = \{p_1, \ldots, p_N\}, \tag{4.1.1}$$

le régime *enfant*

$$E_t \mid P_t = p_i \in \mathring{\mathcal{C}}_i = \left\{e_1^i, \ldots, e_{k_i}^i\right\}, \quad \forall\, i \in \{1, \ldots, N\} \tag{4.1.2}$$

et le régime *producteur*

$$C_t \in \mathcal{C} = \left\{\left(p_1, e_1^1\right), \ldots, \left(p_1, e_{k_1}^1\right), \ldots, \left(p_N, e_1^N\right), \ldots, \left(p_N, e_{k_N}^N\right)\right\}. \tag{4.1.3}$$

Dans le cas qui nous intéresse, nous choisissons à nouveau de simplement numéroter les régimes en fonction de leur position dans l'espace d'états. En l'occurrence, les séries de régimes *producteurs* peuvent prendre leurs valeurs dans l'espace d'états $\mathcal{C} = \{(1,1), (1,2), (2,1), (2,2)\} \equiv \{1, 2, 3, 4\}$, chacune de ces valeurs pointant vers une distribution d'émission.

Spécification 4 : HMM Hiérarchique général

Distribution initiale	$C_1 \sim p(c_1)$
Probabilité de transition	$C_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t-1} = c_{1:t-1}) \sim p(c_t \mid c_{t-1})$
Probabilité d'émission	$X_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t} = c_{1:t}) \sim p(x_t \mid c_t)$

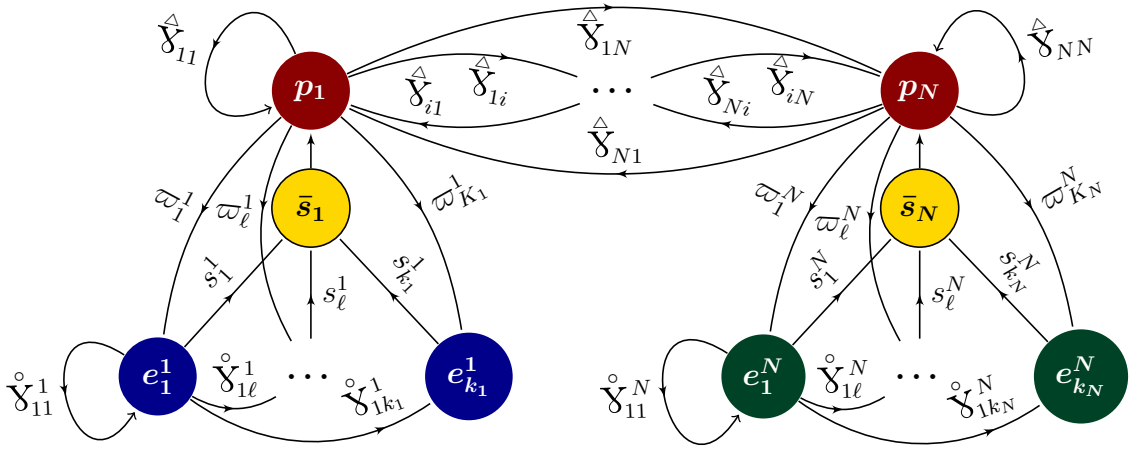
Remarque

Cette spécification ne fait qu'introduire la paire $\{(P_t, E_t)\}_{t\in\mathbb{N}^\star} \equiv \{C_t\}_{t\in\mathbb{N}^\star}$ liée par des relations déterministes. De plus, la construction de ce modèle nécessite le même nombre de paramètres que pour la construction d'un modèle de Markov à variables latentes ayant le même nombre d'états finaux.

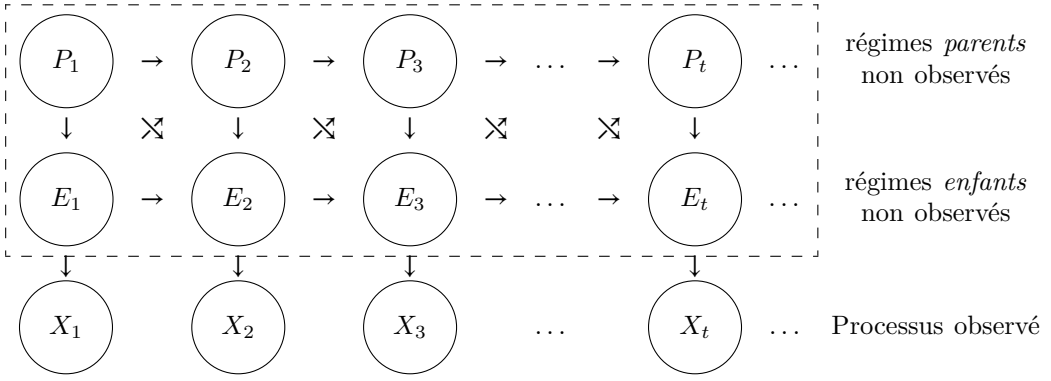
Nouvelle notation associée

$$\overset{\triangle}{\gamma}_{ij} \overset{\text{def}}{=} \mathbb{P}[P_{t+1} = p_j \mid P_t = p_i], \qquad \varpi_\ell^j \overset{\text{def}}{=} \mathbb{P}\Big[E_t = e_\ell^j \;\Big|\; P_t = p_j\Big],$$
$$\overset{\circ}{\gamma}_{\eta\ell}^{\,j} \overset{\text{def}}{=} \mathbb{P}\Big[E_{t+1} = e_\ell^j \;\Big|\; P_{t+1} = p_j, P_t = p_j, E_t = e_\eta^j\Big],$$
$$s_\eta^i \overset{\text{def}}{=} \mathbb{P}\Big[\text{Quitter le régime } \textit{enfant } e_\eta^i \text{ et retourner à son parent } p_i\Big].$$

Représentation graphique du processus latent



Représentation graphique du modèle complet



4.1.1. Architecture de la matrice de transition

Dans un premier temps, rappelons les symboles associés à ce type de structure—ainsi que leurs significations—pour un modèle avec N régimes *parents* et leurs k_i régimes *enfants* associés :

$$\overset{\triangle}{\gamma}_{ij} = \mathbb{P}[\,P_{t+1} = p_j \mid P_t = p_i\,], \tag{4.1.4}$$

$$\varpi_\ell^j = \mathbb{P}\Big[\,E_t = e_\ell^j \;\Big|\; P_t = p_j\,\Big], \tag{4.1.5}$$

$$\overset{\circ}{\gamma}{}_{\eta\ell}^{\,j} = \mathbb{P}\Big[\,E_{t+1} = e_\ell^j \;\Big|\; P_{t+1} = p_j, P_t = p_j, E_t = e_\eta^j\,\Big], \tag{4.1.6}$$

$$s_\ell^i = \mathbb{P}\Big[\text{Quitter le régime } \textit{enfant } e_\ell^i \text{ et retourner à son parent } p_i\Big]. \tag{4.1.7}$$

Par sa structure, le modèle impose à ces probabilités probabilités de transition les contraintes suivantes $\forall\, i \in \{1, \ldots, N\}$:

$$\sum_{j=1}^{N} \overset{\triangle}{\gamma}_{ij} = 1, \tag{4.1.8}$$

$$\sum_{\ell=1}^{k_i} \overset{\circ}{\gamma}{}_{\eta\ell}^{\,i} + s_\eta^i = 1, \quad \forall\, \eta \in \{1, \ldots, k_i\}, \tag{4.1.9}$$

$$\sum_{\ell=1}^{k_i} \varpi_\ell^i = 1. \tag{4.1.10}$$

Ainsi, nous avons besoin dans un premier temps de $N(N-1)$ probabilités pour peupler la chaîne de Markov de la couche supérieure (régimes *parents*)—soit $\Big\{\overset{\triangle}{\gamma}_{ij}\Big\}_{\substack{i\in\{1,\ldots,N\},\\ j\in\{1,\ldots,N-1\}}}$. Pour ce qui est de peupler la couche suivante (régimes *enfants*), le i^{e} parent nécessite $(k_i - 1)$ probabilités pour choisir un de ses k_i enfants—soit $\{\varpi_\ell^i\}_{\ell\in\{1,\ldots,(k_i-1)\}}$—et de k_i^2 probabilités de transition pour peupler la chaîne de Markov de cette branche de la couche subalterne—soit $\Big\{\overset{\circ}{\gamma}{}_{\eta\ell}^{\,i}\Big\}_{\eta,\ell\in\{1,\ldots,k_i\}}$—, les probabilités de choisir le $k_i{}^{\text{e}}$ enfant $(\varpi_{k_i}^i)$ et de sortir d'un régime *enfant* (s_η^i) étant obtenues par soustraction. Ceci représente au total $N(N-1) + \sum_{i=1}^{N}\Big[(k_i - 1) + k_i^2\Big]$ paramètres pour le processus latent des *HHMMs* sous la spécification hiérarchique générale (p.78) à deux couches. Pour le modèle qui nous intéresse (i.e. celui ayant $N = 2$ régimes *parents* avec chacun $k_i = 2$ enfants), au total 12 paramètres libres sont donc nécessaires afin de paramétrer la matrice de transition du processus latent, exactement comme pour un *HMM* classique à 4 régimes tel que présenté à la spécification standard (p.56).

Remarque 4.1.1. Bien que cette architecture ne réduise pas le nombre de paramètres du modèle, il est clair que de nouvelles informations concernant le processus de changement de régime sont obtenues grâce à celle-ci. Cette caractéristique est illustrée clairement à la portion

représentation graphique du processus latent de cette spécification. C'est donc principalement afin de mieux comprendre le processus latent que nous sommes emmenés à considérer la spécification hiérarchique générale (p.78) des modèles de Markov à variables latentes. Quant à la possibilité de paramétrer ces modèles plus parcimonieusement que leurs homologues de la spécification standard (p.56), cette caractéristique d'intérêt sera discutée à la section 4.2 (p.82).

Afin de bien illustrer la construction de la matrice de transition, nous utiliserons une approche par bloc. Nous présenterons d'abord la construction des blocs correspondant à un changement de régime *enfant* mais associé à un même parent, puis des blocs relatifs à la sortie du régime *parent*, la sélection d'un autre parent, puis d'un des enfants de ce dernier. La matrice de transition prendra donc la forme

$$\mathbf{\Gamma} = \begin{bmatrix} \begin{bmatrix} p_1 \to p_1 \end{bmatrix} & \cdots & \begin{bmatrix} p_1 \to p_N \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} p_N \to p_1 \end{bmatrix} & \cdots & \begin{bmatrix} p_N \to p_N \end{bmatrix} \end{bmatrix}. \tag{4.1.11}$$

Intéressons-nous donc dans un premier temps aux transitions dans un même parent, soit les blocs $p_i \to p_i$. En nous fiant à la représentation graphique du processus latent présentée à la spécification hiérarchique générale (p.78), le passage de l'état e_η^i à l'état e_ℓ^i peut se faire en suivant deux parcours. Soit le passage est direct de e_η^i vers e_ℓ^i, ou bien le parcours débute par un retour au i^{e} parent, la sélection de celui-ci à nouveau et le choix du régime e_ℓ^i depuis ce dernier. Nous avons donc les probabilités

$$\begin{aligned} \gamma_{e_\eta^i e_\ell^i} &= \mathbb{P}\Big[E_{t+1} = e_\ell^i \;\Big|\; P_{t+1} = p_i, P_t = p_i, E_t = e_\eta^i\Big] \\ &\quad + \Big(\mathbb{P}\Big[\text{Quitter le régime } \textit{enfant } e_\eta^i \text{ et retourner à son parent } p_i\Big] \\ &\qquad \times \mathbb{P}[P_{t+1} = p_i \mid P_t = p_i] \\ &\qquad \times \mathbb{P}\Big[E_{t+1} = e_\ell^i \;\Big|\; P_{t+1} = p_i\Big]\Big), \\ &= \mathring{\gamma}_{\eta\ell}^{\,i} + s_\eta^i \cdot \mathring{\gamma}_{ii} \cdot \varpi_\ell^i, \quad \begin{cases} \forall\, \eta, \ell \in \{1, \ldots, k_i\} \text{ et} \\ \forall\, i \in \{1, \ldots, N\}. \end{cases} \end{aligned} \tag{4.1.12}$$

Pour ce qui est maintenant des transitions vers un régime *enfant* associé à un parent différent, il suffit de quitter le régime *enfant* actuel, transitionner ensuite vers un autre parent et finalement choisir un des enfants de celui-ci. En termes mathématiques, le passage du régime e_η^i vers le régime e_ℓ^j se fait avec probabilité

$$\begin{aligned}\mathcal{Y}_{e_\eta^i e_\ell^j} &= \Big(\mathbb{P}\Big[\text{Quitter le régime } \textit{enfant } e_\eta^i \text{ et retourner à son parent } p_i\Big] \\ &\qquad \times \mathbb{P}[P_{t+1} = p_j \mid P_t = p_i] \\ &\qquad \times \mathbb{P}\Big[E_{t+1} = e_\ell^j \;\Big|\; P_{t+1} = p_j\Big]\Big), \\ &= s_\eta^i \cdot \overset{\triangle}{\mathcal{Y}}_{ij} \cdot \varpi_\ell^j, \quad \begin{cases} \forall\, \eta \in \{1, \ldots, k_i\}, \\ \forall\, \ell \in \{1, \ldots, k_j\} \text{ et} \\ \forall\, i, j \in \{1, \ldots, N\}. \end{cases} \end{aligned} \tag{4.1.13}$$

Si tous les régimes *parents* ont le même nombre de régimes *enfants* (disons $k_i = k, \forall\, i \in \{1, \ldots, N\}$), la matrice de transition pour les $K = N \cdot k$ régimes *producteurs* peut être obtenue par la formule matricielle

$$\boldsymbol{\Gamma} = \bigoplus_{i=1}^{N} \mathring{\boldsymbol{\Gamma}}_i + \Big[\overset{\triangle}{\boldsymbol{\Gamma}} \otimes (\mathbf{1}_k \mathbf{1}_k{}')\Big] \odot (\mathbf{1}_K \vec{\varpi}) \odot (\mathbf{1}_K \vec{s})', \tag{4.1.14}$$

avec

$$\mathring{\boldsymbol{\Gamma}}_i \overset{\text{def}}{=} \Big(\mathring{\mathcal{Y}}_{\eta\ell}^{\,i}\Big)_{\eta,\ell \in \{1,\ldots,k\}}, \tag{4.1.15}$$

$$\overset{\triangle}{\boldsymbol{\Gamma}} \overset{\text{def}}{=} \Big(\overset{\triangle}{\mathcal{Y}}_{ij}\Big)_{i,j \in \{1,\ldots,N\}}, \tag{4.1.16}$$

$$\vec{\varpi} \overset{\text{def}}{=} (\vec{\varpi}_1, \ldots, \vec{\varpi}_N), \tag{4.1.17}$$

$$\vec{\varpi}_i \overset{\text{def}}{=} \Big(\varpi_1^i, \ldots, \varpi_k^i\Big), \tag{4.1.18}$$

$$\vec{s} \overset{\text{def}}{=} (\vec{s}_1, \ldots, \vec{s}_N), \tag{4.1.19}$$

$$\vec{s}_i \overset{\text{def}}{=} \Big(s_1^i, \ldots, s_k^i\Big) \tag{4.1.20}$$

et où $\bigoplus_{i=1}^{N} \mathring{\boldsymbol{\Gamma}}_i$ correspond à la somme directe des matrices de transition de la couche subalterne, $\otimes$ correspond au produit de Kronecker et $\odot$ est le produit de Hadamard.

Exemple 4.1.1 (Construction de la matrice de transition)**.** *Soit un* HHMM *ayant* 2 *couches cachées structurées comme à la spécification hiérarchique générale* (p. 78) *et où les chaînes de Markov admettent chacune* 2 *régimes. La matrice de transition prend alors la forme*

$$\boldsymbol{\Gamma} = \left[\begin{array}{cc:cc} \left(\overset{\circ}{\gamma}{}^1_{11} + s^1_1 \cdot \overset{\triangle}{\gamma}_{11} \cdot \varpi^1_1\right) & \left(\overset{\circ}{\gamma}{}^1_{12} + s^1_1 \cdot \overset{\triangle}{\gamma}_{11} \cdot \varpi^1_2\right) & \left(s^1_1 \cdot \overset{\triangle}{\gamma}_{12} \cdot \varpi^2_1\right) & \left(s^1_1 \cdot \overset{\triangle}{\gamma}_{12} \cdot \varpi^2_2\right) \\ \left(\overset{\circ}{\gamma}{}^1_{21} + s^1_2 \cdot \overset{\triangle}{\gamma}_{11} \cdot \varpi^1_1\right) & \left(\overset{\circ}{\gamma}{}^1_{22} + s^1_2 \cdot \overset{\triangle}{\gamma}_{11} \cdot \varpi^1_2\right) & \left(s^1_2 \cdot \overset{\triangle}{\gamma}_{12} \cdot \varpi^2_1\right) & \left(s^1_2 \cdot \overset{\triangle}{\gamma}_{12} \cdot \varpi^2_2\right) \\ \hdashline \left(s^2_1 \cdot \overset{\triangle}{\gamma}_{21} \cdot \varpi^1_1\right) & \left(s^2_1 \cdot \overset{\triangle}{\gamma}_{21} \cdot \varpi^1_2\right) & \left(\overset{\circ}{\gamma}{}^2_{11} + s^2_1 \cdot \overset{\triangle}{\gamma}_{22} \cdot \varpi^2_1\right) & \left(\overset{\circ}{\gamma}{}^2_{12} + s^2_1 \cdot \overset{\triangle}{\gamma}_{22} \cdot \varpi^2_2\right) \\ \left(s^2_2 \cdot \overset{\triangle}{\gamma}_{21} \cdot \varpi^1_1\right) & \left(s^2_2 \cdot \overset{\triangle}{\gamma}_{21} \cdot \varpi^1_2\right) & \left(\overset{\circ}{\gamma}{}^2_{12} + s^2_2 \cdot \overset{\triangle}{\gamma}_{22} \cdot \varpi^2_1\right) & \left(\overset{\circ}{\gamma}{}^2_{22} + s^2_2 \cdot \overset{\triangle}{\gamma}_{22} \cdot \varpi^2_2\right) \end{array}\right]. \tag{4.1.21}$$

4.2. Structure simplifiée

Comme nous le mentionnions précédemment, il est aussi possible de structurer un *HHMM* afin de rendre le modèle plus parcimonieux, tout en augmentant la granularité du modèle par rapport à un *HMM* ayant le même nombre de paramètres. La définition d'un *HHMM* simplifié est présentée à la spécification hiérarchique simplifiée (p.83). À nouveau, nous y présentons les représentations graphiques pour un modèle ayant deux couches cachées, mais cette structure peut naturellement être étendue pour avoir un nombre plus élevé de couches.

Sous cette spécification, les variables aléatoires du processus latent d'un *HHMM* avec N régimes *parents* ayant chacun k régimes *enfants* sont telles que, $\forall\, t \in \mathbb{N}^\star$,

$$P_t \in \overset{\triangle}{\mathcal{C}} = \{p_1, \ldots, p_N\}, \tag{4.2.1}$$

$$E_t \in \overset{\circ}{\mathcal{C}} = \{e_1, \ldots, e_k\}. \tag{4.2.2}$$

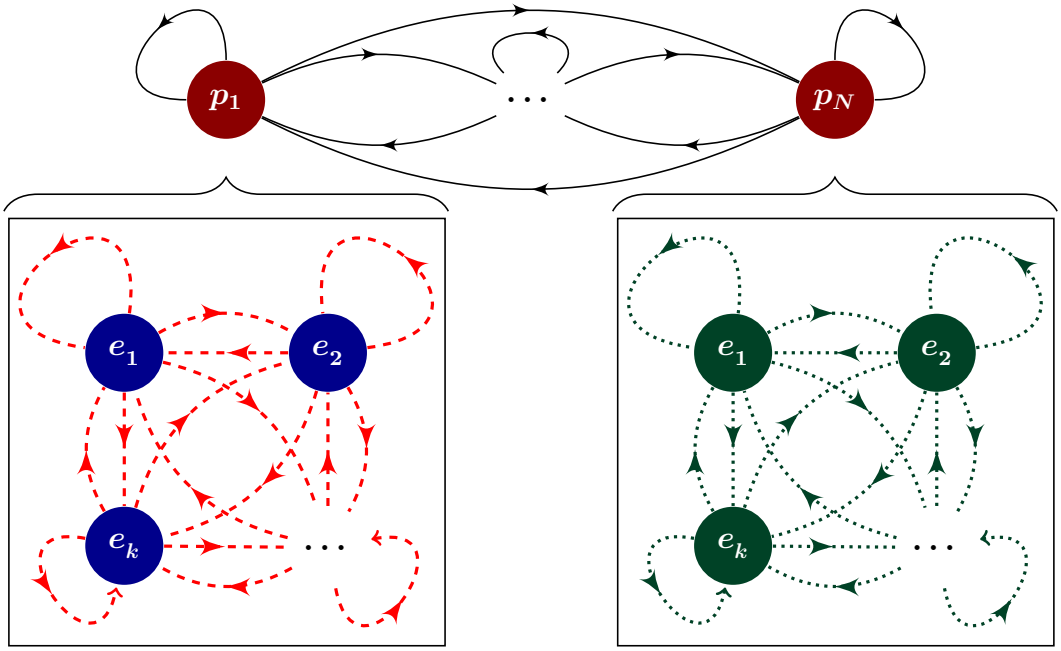
À la différence de la spécification hiérarchique générale (p.78), les branches de la couche subalterne admettent ici le même espace d'états $\overset{\circ}{\mathcal{C}}$. L'espace d'états des régimes *producteurs* est donc $\mathcal{C} = \{(p_1, e_1), \ldots, (p_1, e_k), \ldots, (p_N, e_1), \ldots, (p_N, e_k)\}$, avec $\mathrm{Card}(\mathcal{C}) = k \cdot N$.

Dans le cas qui nous intéresse, les régimes *producteurs* admettent les mêmes combinaisons que dans la spécification hiérarchique générale (p.78), soit $C_t \in \{(1,1),(1,2),(2,1),(2,2)\} \equiv \{1,2,3,4\}$.

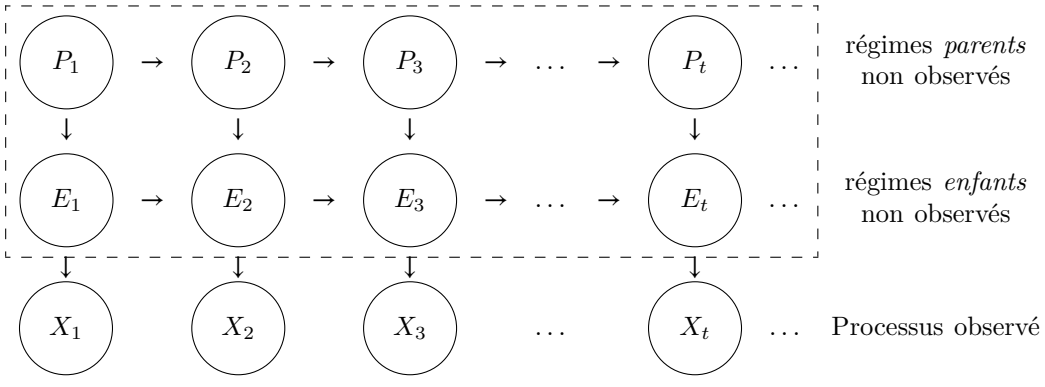
Spécification 5 : HMM Hiérarchique simplifié

Distribution initiale	$C_1 \sim p(c_1)$
Probabilité de transition	$C_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t-1} = c_{1:t-1}) \sim p(c_t \mid c_{t-1})$
Probabilité d'émission	$X_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t} = c_{1:t}) \sim p(x_t \mid c_t)$

Représentation graphique du processus latent



Représentation graphique du modèle complet



4.2.1. Architecture de la matrice de transition

Sous cette spécification simplifiée, chaque couche est indépendante des couches lui étant subalternes. Cependant, le régime émis au temps t par une couche subalterne dépend du régime émis au temps $t-1$ par cette même couche, et du régime émis au temps t par la couche du dessus. Ce ne sont en fait que les probabilités de transition de la couche subalterne qui changent d'un régime *parent* à un autre, l'espace d'états restant fixe.

Les notations utilisées pour la spécification hiérarchique générale (p.78) peuvent à nouveau servir à peupler les chaînes de Markov des couches homologues dans la spécification hiérarchique simplifiée (p.83). Les symboles en question—et leurs interprétations— sont sensiblement les mêmes, mis à part le fait que sous cette spécification les probabilités de transition sur la couche subalterne ne dépendent pas du parent précédent. Nous retenons, $\forall\, p_i, p_j \in \overset{\triangle}{\mathcal{C}}$ et $\forall\, e_\eta, e_\ell \in \mathring{\mathcal{C}}$, les symboles

$$\overset{\triangle}{\mathcal{Y}}_{ij} = \mathbb{P}[\,P_{t+1} = p_j \mid P_t = p_i\,], \tag{4.2.3}$$

$$\mathring{\mathcal{Y}}^{\,i}_{\eta\ell} \overset{\text{hyp}}{=} \mathbb{P}\Big[\,E_{t+1} = e_\ell \;\Big|\; P_{t+1} = p_i, E_t = e_\eta\,\Big], \tag{4.2.4}$$

ainsi que les notations matricielles

$$\overset{\triangle}{\mathbf{\Gamma}} \overset{\text{def}}{=} \Big(\overset{\triangle}{\mathcal{Y}}_{ij}\Big)_{i,j\in\{1,\ldots,N\}}, \tag{4.2.5}$$

$$\mathring{\mathbf{\Gamma}}_i \overset{\text{def}}{=} \Big(\mathring{\mathcal{Y}}^{\,i}_{\eta\ell}\Big)_{\eta,\ell\in\{1,\ldots,k\}}. \tag{4.2.6}$$

Sous cette spécification des *HHMMs*, les contraintes imposées sur les probabilités de transition sont séparées par couche et identiques pour chacune d'entre elles. Ces contraintes se résument en fait à assurer que les matrices de transition $\overset{\triangle}{\mathbf{\Gamma}}$ et $\mathring{\mathbf{\Gamma}}_i, \forall\, i \in \{1,\ldots,N\}$ soient des matrices stochastiques. Sous forme matricielle, nous contraignons donc

$$\overset{\triangle}{\mathbf{\Gamma}} \geq 0, \tag{4.2.7}$$

$$\overset{\triangle}{\mathbf{\Gamma}} \cdot \mathbf{1}_N = \mathbf{1}_N \tag{4.2.8}$$

et

$$\mathring{\mathbf{\Gamma}}_i \geq 0, \tag{4.2.9}$$

$$\mathring{\mathbf{\Gamma}}_i \cdot \mathbf{1}_k = \mathbf{1}_k. \tag{4.2.10}$$

Ainsi, nous avons besoin de $N(N-1)$ paramètres pour la couche supérieure, et de $k(k-1)$ paramètres pour chaque chaîne de la couche subalterne. Ceci correspond donc à un total de $N(N-1) + Nk(k-1)$ paramètres pour le processus latent des *HHMMs* simplifiés à deux couches. Pour le modèle ayant $N = 2$ régimes *parents* admettant chacun $k = 2$ régimes *enfants*, nous avons donc besoin de 6 paramètres pour peupler le processus latent. Cette

valeur est bien en deçà des 12 paramètres requis pour un *HMM* classique à 4 régimes tel que présenté à la spécification standard (p.56).

Pour ce qui est maintenant de calculer les probabilités de transition pour les régimes *producteurs* d'un *HHMM* simplifié, celles-ci sont simplement obtenues en appliquant le théorème de Bayes. Sans perte de généralité, la transition de la paire $\big(p_i, e_\eta\big)$ vers la paire (p_j, e_ℓ) se fait avec probabilité

$$\begin{aligned}
&\mathbb{P}\Big[P_{t+1} = p_j, E_{t+1} = e_\ell \;\Big|\; P_t = p_i, E_t = e_\eta\Big] \\
&\quad \overset{\text{Bayes}}{=\!=} \mathbb{P}\Big[E_{t+1} = e_\ell \;\Big|\; P_{t+1} = p_j, P_t = p_i, E_t = e_\eta\Big] \cdot \mathbb{P}\Big[P_{t+1} = p_j \;\Big|\; P_t = p_i, E_t = e_\eta\Big], \\
&\quad \overset{\text{hyp}}{=\!=} \mathbb{P}\Big[E_{t+1} = e_\ell \;\Big|\; P_{t+1} = p_j, E_t = e_\eta\Big] \cdot \mathbb{P}[P_{t+1} = p_j \mid P_t = p_i], \\
&\quad = \overset{\circ}{\gamma}{}^{j}_{\eta\ell} \cdot \overset{\triangle}{\gamma}_{ij}, \begin{cases} \forall\, p_i, p_j \in \overset{\triangle}{\mathcal{C}} \text{ et} \\ \forall\, e_\eta, e_\ell \in \overset{\circ}{\mathcal{C}}. \end{cases}
\end{aligned}$$

Afin de construire la matrice de transition sous cette spécification, nous nous intéressons au prochain régime *parent*, plutôt qu'au régime *parent* actuel tel que c'était le cas précédemment. Nous utiliserons à cet effet le symbole $\nrightarrow_j$ pour dénoter les colonnes de $\mathbf{\Gamma}$ correspondantes à une transition vers le parent p_j. Ces matrices sont obtenues en prenant le produit de Kronecker de la j^{e} colonne de la matrice $\overset{\triangle}{\mathbf{\Gamma}}$ et de la matrice $\overset{\circ}{\mathbf{\Gamma}}_j$. Ainsi, nous avons que

$$\nrightarrow_j = \overset{\triangle}{\gamma}_{\star j} \otimes \overset{\circ}{\mathbf{\Gamma}}_j, \quad \forall\, j \in \{1, \ldots, N\}, \tag{4.2.11}$$

où $\overset{\triangle}{\gamma}_{\star j}$ désigne la j^{e} colonne de la matrice $\overset{\triangle}{\mathbf{\Gamma}}$. La matrice de transition correspondante aux régimes *producteurs* est alors

$$\mathbf{\Gamma} = \left[\begin{array}{c|c|c} & & \\ \nrightarrow_1 & \cdots & \nrightarrow_N \\ & & \end{array}\right] \tag{4.2.12}$$

avec les $\nrightarrow_j$ de dimensions $(Nk \times k)$, $\forall\, j \in \{1, \ldots, N\}$.

Exemple 4.2.1 (Construction de la matrice de transition)**.** *Soit un* HHMM *ayant* 2 *couches cachées structurées comme à la spécification hiérarchique simplifiée (p.83) et où les chaînes de Markov admettent chacune* 2 *régimes. La matrice de transition prend alors la forme*

$$\mathbf{\Gamma} = \left[\begin{array}{cc:cc} \big(\overset{\triangle}{\gamma}_{11} \cdot \overset{\circ}{\gamma}{}^{1}_{11}\big) & \big(\overset{\triangle}{\gamma}_{11} \cdot \overset{\circ}{\gamma}{}^{1}_{12}\big) & \big(\overset{\triangle}{\gamma}_{12} \cdot \overset{\circ}{\gamma}{}^{2}_{11}\big) & \big(\overset{\triangle}{\gamma}_{12} \cdot \overset{\circ}{\gamma}{}^{2}_{12}\big) \\ \big(\overset{\triangle}{\gamma}_{11} \cdot \overset{\circ}{\gamma}{}^{1}_{21}\big) & \big(\overset{\triangle}{\gamma}_{11} \cdot \overset{\circ}{\gamma}{}^{1}_{22}\big) & \big(\overset{\triangle}{\gamma}_{12} \cdot \overset{\circ}{\gamma}{}^{2}_{21}\big) & \big(\overset{\triangle}{\gamma}_{12} \cdot \overset{\circ}{\gamma}{}^{2}_{22}\big) \\ \hdashline \big(\overset{\triangle}{\gamma}_{21} \cdot \overset{\circ}{\gamma}{}^{1}_{11}\big) & \big(\overset{\triangle}{\gamma}_{21} \cdot \overset{\circ}{\gamma}{}^{1}_{12}\big) & \big(\overset{\triangle}{\gamma}_{22} \cdot \overset{\circ}{\gamma}{}^{2}_{11}\big) & \big(\overset{\triangle}{\gamma}_{22} \cdot \overset{\circ}{\gamma}{}^{2}_{12}\big) \\ \big(\overset{\triangle}{\gamma}_{21} \cdot \overset{\circ}{\gamma}{}^{1}_{21}\big) & \big(\overset{\triangle}{\gamma}_{21} \cdot \overset{\circ}{\gamma}{}^{1}_{22}\big) & \big(\overset{\triangle}{\gamma}_{22} \cdot \overset{\circ}{\gamma}{}^{2}_{21}\big) & \big(\overset{\triangle}{\gamma}_{22} \cdot \overset{\circ}{\gamma}{}^{2}_{22}\big) \end{array}\right] \tag{4.2.13}$$

Remarque 4.2.1 (Stationnarité). Il est intéressant de souligner ici que si les chaînes sur toutes les couches sont ergodiques, la chaîne de Markov des régimes *producteurs* résultante est nécessairement ergodique et admet donc une distribution stationnaire unique.

Chapitre 5

Extensions non-homogènes

À la remarque 1.1.1 (p.37) du chapitre 1, nous mentionnions l'inaptitude des *HMMs* à reproduire le fait sylisé (2), en particulier la lente décroissance de l'autocorrélation dans les processus $\{\epsilon_t^2\}$ et $\{|\epsilon_t|\}$ [1] qui s'apparente à un processus ayant une longue mémoire.

Selon Bulla et Bulla [2006], le fait que le temps de séjour dans un état d'une chaîne de Markov d'ordre un soit par définition géométrique limite la capacité des *HMMs* standards à modéliser la longue mémoire que certaines séries semblent afficher, et par extension leur capacité à reproduire le fait stylisé (2). Ces auteurs proposent à cet effet l'utilisation de modèles semi-markoviens permettant de relaxer l'hypothèse de perte de mémoire des *HMMs*. Cette approche permet donc de donner au modèle une plus longue mémoire par construction.

Dans leurs travaux, Calvet et J. Fisher [2004] proposent plutôt l'utilisation de modèles ayant une très haute dimensionalité dans la chaîne de Markov, soit les *Markov-switching multifractal*—abbréviés *MSM*. Ils démontrent que ces modèles ont la faculté de reproduire une grande persistance dans la volatilité des taux de change en reproduisant artificiellement des effets semi-markoviens, augmentant pour ainsi dire la mémoire du processus.

Dans l'article « Stylized Facts of Daily Return Series and the Hidden Markov Model », Rydén, Teräsvirta et Åsbrink [1998] présentent une étude de l'adéquation des *HMMs* à reproduire les faits stylisés sur la série de l'indice S&P 500 du 3 janvier 1929 au 30 avril 1991. Ces derniers séparent l'échantillon en 10 sous-séries d'environ 1700 observations qui seront modélisées séparément par 10 modèles distincts ayant la même structure fondamentale. En analysant les paramètres estimés pour les différents modèles, ils en arrivent à une observation très intéressante liée à l'enjeu de la mémoire du processus. Les auteurs observent que les modèles entraînés sur ces diverses sous-séries présentent parfois des paramètres très différents—bien que les modèles sur les sous-séries contigues soient généralement plus semblables. Cette observation semble suggérer que la relaxation de l'hypothèse d'homogénéité puissent corriger cette lacune des *HMMs* standard.

1. Rydén, Teräsvirta et Åsbrink [1998].

Bien que relativement peu de recherche ait à ce jour été dédiée à l'étude des structures non-homogènes pour la matrice de transition des *HMMs*, cette relaxation trouve multiples justifications dans divers phénomènes naturels. Nous savons d'ailleurs depuis plusieurs années que l'hypothèse d'homogénéité est très contraignante. Indépendamment les uns des autres, les travaux de Diebold, Lee et Weinbach [1994] et Filardo [1994] démontraient que l'hypothèse d'homogénéité est bien trop contraignante dans une grande quantité de contextes empiriques. Plus spécifiquement par rapport au niveau du S&P 500, Nystrup, Madsen et Lindström [2017] démontrent, suite aux découvertes de Rydén, Teräsvirta et Åsbrink [1998] et Bulla [2011], que la valeur des paramètres liés aux changements de régime varient énormément sur la durée de la série chronologique des log-rendements quotidiens sur l'indice S&P 500. L'utilisation de *HMMs* non-homogènes se justifie donc par des études empiriques, et se démarque par sa capacité à admettre un temps de séjour non-géométrique et à incorporer une forme de dépendance à long-terme, lorsque la structure choisie pour la matrice $\mathbf{\Gamma}_t$ le permet.

Afin d'étudier cette relaxation de l'hypothèse d'homogénéité, nous considérerons des modèles provenant des deux classifications proposées par Cox et al. [1981] pour les modèles dont les paramètres varient dans le temps. La première de ces catégories, soit les modèles dits *observation-driven*, inclue tous les modèles dont les paramètres sont des fonctions déterministes de variables observables. La seconde catégorie, soit les modèles dits *parameter-driven*, réfère aux modèles dont les paramètres varient selon un processus non-observable.

Nous nous intéresserons donc, dans un premier temps, aux structures utilisant des variables explicatives observables—potentiellement des termes précédents de la variable dépendante—afin de permettre à la matrice $\mathbf{\Gamma}_t$ de varier dans le temps. Ces structures incluent, entre autres, celles proposées par Diebold, Lee et Weinbach [1994], Hamilton [1994], Filardo [1994], Perez-Quiros et Timmermann [2000] et Bazzi, Blasques, Koopman et Lucas [2017]. Sous ces structures, la log-vraisemblance peut-être obtenue analytiquement par le filtre d'Hamilton, permettant une estimation des paramètres sans recours à la simulation.

Puisque les modèles de la seconde catégorie incluent des variables ne pouvant être observées—et devant donc être inférées—, ces derniers requièrent habituellement plus d'ajustements à la structure du modèle complet. Nous restreindrons donc notre étude des modèles de cette catégorie à ceux admettant une dépendance au nombre de pas déjà passés dans l'état actuel (e.g. Durland et McCurdy [1994], Lam [1997], Kim et Nelson [1998] et Maheu et McCurdy [2000a ; 2000b]). Nous présenterons en particulier le modèle de Markov à variables latentes avec dépendance à la durée proposé par Maheu et McCurdy [2000a ; 2000b]. Il sera nécessaire pour ce modèle d'ajuster la paramétrisation de la chaîne et de mettre à jour le filtre d'Hamilton afin de construire analytiquement la vraisemblance.

5.1. Transitions dépendantes de variables observables (*Observation-driven models*)

Pour les modèles de cette catégorie, les probabilités de transition dépendent de variables observables au temps t, lesquelles seront contenues dans un vecteur ligne noté $\vec{Y}_t$. Il est possible, par exemple, d'utiliser des valeurs passées de la série à l'étude soit $Y_t = X_t$. D'ailleurs, nous avons déjà présenté un modèle ayant ce type de structure à la spécification généralisée (p.58) introduite au chapitre 3 et utilisant la série à l'étude $\{X_t\}_{t\in\mathbb{N}^\star}$ comme variable explicative dans la construction de $\mathbf{\Gamma}_t$. Ainsi, nous savons déjà que ces structures permettent l'utilisation de l'algorithme 1 : Filtre d'Hamilton (p.68) et nous pourrons donc obtenir la log-vraisemblance du modèle analytiquement. Un autre avantage des modèles de cette catégorie est qu'ils ne viennent pas modifier l'espace d'états de la chaîne de Markov. Les approches présentées dans cette section ne font en fait que spécifier la forme que prennent les termes de la matrice de transition dans la spécification généralisée (p.58).

Dans cette section, nous présenterons d'abord la forme générale que prend la relation entre les variables explicatives et les probabilités de transition, puis deux formes particulières et quelques résultats tirés de la littérature.

5.1.1. Architecture de la matrice de transition

Dans un *HMM* non-homogène avec transitions dépendantes de variables observables, les probabilités de transition prennent habituellement, $\forall\, t \in \mathbb{N}^\star$ et $\forall\, i \in \{1,\ldots,K\}$, la forme

$$\Upsilon_{ij}(t) \stackrel{\text{def}}{=} \mathbb{P}[C_{t+1} = j \mid C_t = i, \vec{Y}_t = \vec{y}_t], \tag{5.1.1}$$

$$= \delta_{ij} + (1 - 2\delta_{ij}) \cdot \frac{\exp\{f_{ij,t}\}}{1 + \sum_{\ell=1}^{K-1} \exp\{f_{i\ell,t}\}}, \quad \forall\, j \in \{1,\ldots,K-1\} \tag{5.1.2}$$

et

$$\Upsilon_{iK}(t) = 1 - \sum_{j=1}^{K-1} \Upsilon_{ij}(t), \tag{5.1.3}$$

où les $\delta_{ij} \in [0, 1/2]$ sont des hyperparamètres et les $f_{ij,t}$ sont des fonctions de variables observables contenues dans le vecteur $\vec{y}_t$. Puisque la fonction exponentielle est toujours non-négative et que les probabilités sont normalisées sur les lignes de $\mathbf{\Gamma}_t$ par définition, cette forme pour les probabilités $\Upsilon_{ij}(t)$ permet que la fonction $f_{i\ell,t}$ ne soit pas contrainte $\forall\, i, j \in \mathcal{C}$ et $t \in \mathbb{N}^\star$.

Dans cette forme générale, les hyperparamètres δ_{ij} permettent de contraindre les probabilités de sorte que $\Upsilon_{ij}(t) \in [\delta_{ij}, 1 - \delta_{ij}]$, $\forall\, t \in \mathbb{N}^\star$. La valeur de ces hyperparamètres pourrait

être choisie afin de limiter l'intervalle dans lequel peuvent se trouver les probabilités de transition de la chaîne cachée en fonction de considérations propres au domaine d'utilisation. Dans la majorité des cas—y compris l'application présentée au chapitre 6 (p.105)—, ces hyperparamètres seront fixés à zéro.

5.1.2. Formes pour les fonctions $f_{ij,t}$

5.1.2.1. Combinaison linéaire de variables observables

L'approche proposée par Diebold, Lee et Weinbach [1994] construit sur les travaux de Lee [1991] et donne à la fonction $f_{ij,t}$ la forme

$$f_{ij,t} = \vec{y}_t \boldsymbol{\beta}_{ij}, \tag{5.1.4}$$

où $\vec{y}_t = \begin{bmatrix} 1 & y_{t,1} & \cdots & y_{t,q} \end{bmatrix}$ est un vecteur ligne contenant $(q+1) \in \mathbb{N}^\star$ variables explicatives observables et $\boldsymbol{\beta}_{ij} = \begin{bmatrix} \beta_{ij}^0 & \beta_{ij}^1 & \cdots & \beta_{ij}^q \end{bmatrix}'$ est un vecteur colonne contenant les coefficients associés à ces variables. Afin de les différencier des autres modèles de ce chapitre, nous noterons ces modèles HMM-LOGIT(K), où K correspond au nombre de régimes de la chaîne de Markov.

Puisque la forme de la matrice de transition ne change pas, il y a, dans un HMM-LOGIT(K), $(K-1)$ probabilités libres par ligne de la matrice $\boldsymbol{\Gamma}_t$. Cette paramétrisation de la matrice de transition nécessitant $(q+1)$ coefficients pour chacune de ces probabilités libres, $(q+1)K(K-1)$ paramètres sont au total nécessaires afin de peupler la chaîne de Markov.

Sous cette spécification de $f_{ij,t}$, les coefficients β_{ij}^0 font en sorte que ce modèle ne soit qu'une généralisation de la spécification standard (p.56). En effet, si nous avions $\beta_{ij}^\ell = 0$ $\forall\, \ell \in \{1, \ldots, q\}$ et $\forall\, i, j \in \mathcal{C}$, alors les probabilités de transition seraient indépendantes du temps t et ce modèle serait équivalent à un *HMM* standard.

Dans leur article à vocation principalement méthodologique, Diebold, Lee et Weinbach [1994] montrent, en utilisant des échantillons simulés avec un *HMM* non-homogène, que leur modèle converge relativement rapidement vers la vraie log-vraisemblance de l'échantillon, avec presque tous les paramètres près de leurs vraies valeurs. Ils montrent de plus que leur modèle performe deux fois mieux qu'un *HMM* homogène lorsque la distribution utilisée pour simuler les données était non-homogène, avec une erreur moyenne au carré pour l'inférence des états cachés [2] (*mean squared state extraction error*) de 0.13 contre 0.27.

2. Voir tableau 5.2.

Ces résultats sont très semblables à ceux obtenues par Filardo [1994]. Ce dernier suggère l'utilisation d'indicateurs économiques [3] comme variables explicatives pour la modélisation du produit intérieur brut de la reserve fédérale (*IP index*) avec un HMM-LOGIT($K = 2$). Bien que le modèle résultant ne soit pas supérieur en tout point aux modèles alternatifs présentés dans cet article, il démontre que celui-ci décelle généralement mieux les dynamiques changeantes au gré des cycles économiques. Il observe en particulier que la persistance des régimes varient selon le contexte économique. Cette caractéristique pourrait permettre à ces modèles de mouler plus étroitement le comportement de la série à l'étude, et potentiellement de reproduire le faits stylisé (2).

Les résultats rapportés par Liu, Margaritis et Wang [2012] soutiennent de plus l'idée que l'inclusion de variables observables dans une dynamique du type proposée par Diebold, Lee et Weinbach puisse permettre d'augmenter la qualité des prévisions issues du modèle. L'un de leurs modèles à deux régimes incorpore la largeur moyenne de l'intervalle de prix hebdomadaire dans les 6 mois précédents comme regresseur dans la matrice de transition. En terme de performance à l'intérieur de l'échantillon d'entraînement (*in-sample*), ce modèle performe mieux que les GARCH$(1,1)$, T-GARCH$(1,1)$, EGARCH$(1,1)$ et MSGARCH$(1,1)$ sans régresseurs pour la prévision de la volatilité un pas en avant, et n'est battu que par le EGARCH$(1,1)$ en terme de *AIC* et de *BIC*[4]. La performance de ce modèle à l'extérieur de l'échantillon d'entraînement (*out-of-sample*) est quant à elle supérieure à tous les autres modèles pour l'erreur de prévision moyenne au carré (*mean squared forecast error*, *MSFE*) et en valeur absolue (*mean absolute forecast error*, *MAFE*) [5]. Ces modèles se démarquent par ailleurs en ayant les plus petites valeurs de *MSFE* et *MAFE* pour la prévision de la moyenne à l'intérieur et à l'extérieur de l'échantillon.

Tableau 5.1 – Critères de sélection de modèles basés sur la log-vraisemblance, le nombre de paramètres et la taille de l'échantillon

Akaike Information Criterion	$: \mathrm{AIC}_K = -2\ell_T\left(\widehat{\boldsymbol{\theta}}\right) + 2D_K$
Akaike Information Criterion with small sample correction	$: \mathrm{AICc}_K = \mathrm{AIC}_K + \frac{2D_K^2+2D_K}{T-D_K-1}$
Bayesian Information Criterion	$: \mathrm{BIC}_K = -2\ell_T\left(\widehat{\boldsymbol{\theta}}\right) + \log(T)D_K$

avec
— T : la taille de l'échantillon,
— D_K : le nombre de paramètres libres d'un modèle à K régimes

3. e.g. des indices composites (CLI, XLI, SP), l'indice de diffusion du CLI (DFI), *Term Premium* (TP) et le *Federal Funds rate* (FF).
4. Voir tableau 5.1.
5. Voir tableau 5.2.

Tableau 5.2 – Critères de sélection de modèles basés sur l'erreur de prévision

Mean Absolute Error	$: MAE$	$= \frac{1}{T}\sum_{t=1}^{T} \lvert y_t - \hat{y}_t \rvert$
Mean Squared Error	$: MSE$	$= \frac{1}{T}\sum_{t=1}^{T} (y_t - \hat{y}_t)^2$
Mean Absolute Forecast Error	$: MAFE$	$= \frac{1}{T}\sum_{t=1}^{T} \lvert x_t - \mu_t \rvert$
Mean Squared Forecast Error	$: MSFE$	$= \frac{1}{T}\sum_{t=1}^{T} (x_t - \mu_t)^2$
Mean Absolute Standardized Forecast Error	$: MASFE$	$= \frac{1}{T}\sum_{t=1}^{T} \left\lvert \frac{x_t - \mu_t}{\sigma_t} \right\rvert$
Mean Squared Standardized Forecast Error	$: MSSFE$	$= \frac{1}{T}\sum_{t=1}^{T} \left(\frac{x_t - \mu_t}{\sigma_t} \right)^2$

avec
— y_t une série quelconque et $\hat{y}_t$ la série des valeurs inférées de cette dernière,
— $\mu_t = \mathbb{E}[X_t \mid \boldsymbol{\mathcal{F}}_{t-1}]$,
— $\sigma_t^2 = \mathbb{V}ar\Big[(X_t - \mu_t)^2 \;\Big|\; \boldsymbol{\mathcal{F}}_{t-1}\Big]$.
L'erreur moyenne au carré pour l'inférence des états cachés (*mean squared state extraction error* ou *MSSEE*) n'est autre chose qu'un cas particulier de la *MSE* sur les paires $\left(C_t, \hat{C}_t\right)$, où $\hat{C}_t$ est la valeur inférée de l'état caché au temps t.

Plus récemment, les résultats obtenus par Tan, Koh, Ng et Ng [2020] concernant la modélisation des dynamiques de la volatilité des rendements de la série BTC (*Bitcoin*) vont dans la même direction. Ces derniers incorporent au processus de changement de régime de la variance le logarithme du volume d'échange quotidien du *Bitcoin* ou la quantité de recherches faites quotidiennement sur *Google* dans une dynamique de type HMM-LOGIT(K). Ils démontrent que, dans ce contexte, les modèles résultants—qu'ils notent TV-MSGARCH (*Time-varying Markov-switching* GARCH)—sont généralement plus performants que les modèles GARCH et MSGARCH en terme de log-vraisemblance, de *AIC*, de *AICc*[6], de *BIC* et d'erreur de prévision à l'intérieur et à l'extérieur de l'échantillon d'entraînement du modèle.

5.1.2.2. Dynamique généralisée autorégressive du score de la densité prédictive

Discutés d'abord dans Creal, Koopman et Lucas [2008 ; 2011], puis proposés dans Creal, Koopman et Lucas [2013], les modèles avec dynamiques autorégressives des scores généralisées (*Generalized Autoregressive Score models*)—abbréviés *GAS*—incorporent le score de la densité prédictive mis à l'échelle et la valeur précédente des paramètres au processus de variation des paramètres du modèle. C'est par la suite dans les travaux de Bazzi, Blasques, Koopman et Lucas [2017] que cette approche est utilisée dans le contexte

6. Voir tableau 5.1.

des modèles de Markov à variables latentes, donnant lieu à des modèles que nous noterons HMM-GAS(K), où K correspond au nombre de régimes de la chaîne de Markov.

Dans leur article, ces derniers auteurs proposent une formule matricielle pour construire les fonctions $f_{ij,t}, \forall\, i, j \in \mathcal{C}$. Ils rassemblent ces fonctions dans un vecteur noté $\boldsymbol{f}_t \in \mathbb{R}^{K(K-1)}$ qui est obtenu par la relation de type autorégressive

$$\boldsymbol{f}_{t+1} = \boldsymbol{\omega} + \mathbf{A}\mathbf{s}_t + \mathbf{B}\boldsymbol{f}_t, \tag{5.1.5}$$

où

$$\mathbf{s}_t = \mathbf{S}_t \boldsymbol{\nabla}_t, \tag{5.1.6}$$

$$\boldsymbol{\nabla}_t = \frac{\partial}{\partial \boldsymbol{f}_t} \mathbb{P}[X_t \mid \boldsymbol{\mathcal{F}}_{t-1}, \boldsymbol{f}_t, \boldsymbol{\theta}] \tag{5.1.7}$$

et dans laquelle $\boldsymbol{\omega} \in \mathbb{R}^{K(K-1)}$ est un vecteur de constantes, $\mathbf{A}$ et $\mathbf{B}$ sont des matrices $K(K-1) \times K(K-1)$—habituellement diagonales—de coefficients constants et $\mathbf{S}_t$ est une matrice de mise à l'échelle pour le score qui varie dans le temps. Puisque celle-ci tient compte de la courbure de $\boldsymbol{\nabla}_t$ par rapport à $\boldsymbol{f}_t$, les auteurs proposent pour $\mathbf{S}_t$ d'utiliser la racine carrée de l'inverse de la matrice d'information de Fisher [7].

Cette spécification pour les fonctions $f_{ij,t}$ nécessite donc $2K(K-1)\Big(K(K-1)+1\Big)$ paramètres lorsque les matrices $\mathbf{A}$ et $\mathbf{B}$ sont complètes et $4K(K-1)$ paramètres lorsqu'elles sont diagonales.

Si, parmi les paramètres constants propres à cette spécification de $f_{ij,t}$, seul le vecteur $\boldsymbol{\omega}$ contenait des éléments non-nuls, alors les probabilités de transition seraient indépendantes du temps t et ce modèle serait équivalent à un *HMM* standard.

Afin de justifier l'utilisation de dynamiques *GASs* dans les *HMMs*, Bazzi, Blasques, Koopman et Lucas démontrent l'efficacité du modèle HMM-GAS(K) à la fois sur des données simulées par une approche Monte-Carlo et sur la série du niveau de la production industrielle totale américaine (*INDPRO*).

À cet effet, ils simulent des données à partir de six spécifications différentes pour les probabilités $\Upsilon_{11}(t)$ et $\Upsilon_{22}(t)$ (voir tableau 5.3 (p.94))—dont aucune n'incorpore de dynamiques *GAS*. Sur ces échantillons, leurs résultats indiquent que le HMM-GAS($K=2$) performe en moyenne mieux en terme de log-vraisemblance, de *AICc* et de *BIC* qu'un *HMM* homogène et qu'un HMM-LOGIT($K=2$). Plus précisément, le HMM-GAS($K=2$) performe beaucoup mieux en terme de *AICc* et *BIC* que le modèle homogène (constant), et n'est supplanté par ce dernier que lorsque les données sont générées par un processus homogène ou affichant une cassure unique au centre de l'échantillon. Il en va de même pour l'erreur moyenne au carré (*MSE*) et en valeur absolue (*MAE*) sur l'estimation des probabilités filtrées du processus.

7. Une définition est présentée dans l'appendice externe.

Étonnemment, pour leur HMM-GAS($K = 2$), ces statistiques semblent de plus décroître lorsque la taille de l'échantillon augmente, ce qui n'est pas le cas pour les autres modèles.

Les résultats de leur étude indiquent de surcroît que le HMM-GAS($K = 2$) perfome mieux en terme de *MAFE*, *MSFE*, *MASFE* et *MSSFE*[8] sur toutes les séries générées par un processus dont les transitions sont non-homogènes. Étonnamment, le HMM-GAS($K = 2$) performe mieux en terme d'erreurs de prédiction non-standardisées que le HMM-LOGIT($K = 2$), même lorsque c'est ce dernier qui avait simulé les données.

Leur étude par simulation semble donc démontrer que le HMM-GAS($K = 2$) est le modèle qui performe le mieux lorsqu'il ne correspond pas au processus de génération des observations. En d'autres mots, leurs résultats indiquent que c'est ce modèle qui s'ajuste le plus facilement lorsqu'il n'est pas correctement spécifié, parfois éclipsé mais seulement par le modèle étant correctement spécifié.

Tableau 5.3 – Spécifications de $\Upsilon_{11}(t)$ et $\Upsilon_{22}(t)$ pour simulations

	Modèle	$\Upsilon_{11}(t)$	$\Upsilon_{22}(t)$
1.	Homogène	0.95	0.85
2.	Sin lent	$0.5 + 0.45\sin(4\pi t/T)$	$0.5 - 0.45\sin(4\pi t/T)$
3.	Sin rapide	$0.5 + 0.45\sin(20\pi t/T)$	$0.5 + 0.45\sin(20\pi t/T)$
4.	Cassure	$0.2 \cdot \mathbb{1}_{\{t<T/2\}} + 0.8 \cdot \mathbb{1}_{\{t\geq T/2\}}$	$0.8 \cdot \mathbb{1}_{\{t<T/2\}} + 0.2 \cdot \mathbb{1}_{\{t\geq T/2\}}$
5.	Petite cassure	$0.4 \cdot \mathbb{1}_{\{t<T/2\}} + 0.6 \cdot \mathbb{1}_{\{t\geq T/2\}}$	$0.6 \cdot \mathbb{1}_{\{t<T/2\}} + 0.4 \cdot \mathbb{1}_{\{t\geq T/2\}}$
6.	Autoregressif (décalée)	$(1 + \exp\{-1.2x_{t-1}\})^{-1}$	$(1 + \exp\{0.3x_{t-1}\})^{-1}$
7.	Exogène	$(1 + \exp\{0.7y_{t-1}\})^{-1}$	$(1 + \exp\{-0.7y_{t-1}\})^{-1}$

où $y_t = \phi y_{t-1} + \sqrt{1-\phi^2}\nu_t$, avec $\nu_t \sim \mathcal{N}(0,1)$ et $\phi = 0.99$.
Tableau tiré de Bazzi, Blasques, Koopman et Lucas [2017](p. 465).

Tel que mentionné précédemment, leur étude se conclut par une comparaison de la performance de trois modèles sur la série de la production industrielle américaine, soit à nouveau un *HMM* homogène, un HMM-LOGIT(K) et leur HMM-GAS(K). Les trois modèles considérés admettent 3 régimes pour la moyenne—à laquelle ils donnent une forme autorégressive—et 2 régimes pour la volatilité. Ces deux processus ayant chacun une matrice de transition à part entière, la matrice de transition du modèle complet est obtenue par le produit de Kronecker de celles-ci, à l'instar de Doornik [2013]. Il est important de souligner que dans leur définition de ces modèles, seul le processus de changement de régime pour la variance admet une structure non-homogène. De plus, afin de ne pas contraindre les probabilités de transition, tous les coefficents δ_{ij} sont fixés à zéro pour tous les modèles dont les paramètres varient

8. Voir tableau 5.2.

dans le temps. Leurs résultats indiquent que leur modèle est le plus performant sur cette série, capturant plus efficacement les caractéristiques variables du processus de changement de régime. En effet, leur modèle affiche une grande persistance dans les deux régimes, les entrées de la matrice diagonale **A** étant positives, la matrice diagonale **B** étant presque égal à la matrice identité et la variation des probabilités de rester dans un état coïncidant avec les caractéristiques observées au préalable sur la série.

Bazzi, Blasques, Koopman et Lucas soutiennent finalement, à la lumière de ces résultats, que lorsque l'on suspecte le comportement de la série à l'étude d'être non-homogène sur l'intervalle étudié, les HMM-GAS(K) se présentent comme de bons modèles de référence.

5.2. Transitions dépendantes de variables non-observables (*Parameter-driven models*)

Une autre méthode permettant de faire varier les probabilités de transition dans le temps est d'inclure une autre variable latente dans le processus de changement de régime. Les modèles de cette classe ont malheureusement tendance à être plus complexes à estimer car la variable n'est pas observable au temps t. Il existe cependant des paramétrisations permettant d'inclure une telle variable sans que le modèle ne s'en voit démesurément complexifié. Il faudra tout de même ajuster le filtre d'Hamilton afin qu'il puisse être utilisé dans ce contexte.

Afin d'introduire cette catégorie de modèles, nous étudierons les modèles de Markov à variables latentes avec dépendance à la durée.

5.2.1. *Duration-dependent Markov switching - DDMS*

Les modèles de Markov à variables latentes avec dépendance à la durée (de l'anglais *Duration-Dependent Markov Switching models* et abbréviés *DDMS*)—Durland et McCurdy [1994], Lam [1997], Maheu et McCurdy [2000a ; 2000b]—introduisent dans la matrice de transition une dépendance au nombre de pas consécutifs dans un même état—la durée.

Définition 5.2.1 (Durée (*Duration*))**.** *Soit D_t, la durée de temps—nombre de pas consécutifs—passée dans l'état de la chaîne au temps t. Cette variable est définie par la relation*

$$D_t \stackrel{\text{def}}{=} \min\Big(D_{t-1}\mathbb{1}_{\{C_t=C_{t-1}\}} + 1, \tau\Big) \in \{1, \dots, \tau\}, \tag{5.2.1}$$

où τ est un hyperparamètre permettant de borner supérieurement l'effet du temps passé dans un état en limitant la mémoire du processus.

Remarque 5.2.1. Il est possible d'obtenir la valeur optimale de τ en utilisant des algorithmes de recherches d'hyperparamètres. Maheu et McCurdy [2000a] choisissent pour leur part d'utiliser l'algorithme *grid-search* afin d'obtenir la valeur de τ qui maximise la log-vraisemblance. Nous présenterons pour notre part ces modèles pour un τ quelconque, et référons les intéressé.e.s à Bergstra et Bengio [2012] pour une étude de l'efficacité relative de trois de ces algorithmes, dont *grid-search*.

Jusqu'ici, nous avions considéré exclusivement des chaînes de Markov d'ordre un, c'est-à-dire des chaînes dans lesquelles la variable latente C_t est indépendante de $C_{1:t-2}$ conditionnellement à C_{t-1} (i.e. $\mathbb{P}[C_t \mid C_{1:t-1}] = \mathbb{P}[C_t \mid C_{t-1}]$). Cependant, l'inclusion d'une dépendance à la durée a pour effet d'augmenter l'ordre de la chaîne de Markov ; une chaîne d'ordre τ étant telle que $\mathbb{P}[C_t \mid C_{1:t-1}] = \mathbb{P}[C_t \mid C_{t-\tau:t-1}]$. Cette dépendance de plus grand ordre dans la chaîne de Markov complique techniquement l'estimation des paramètres. En revanche, Maheu et McCurdy [2000a](p. 111-112) soulignent qu'il est possible de reformuler ce modèle comme une chaîne de Markov d'ordre un et proposent une mise-à-jour du filtre d'Hamilton adaptée à cette reparamétrisation.

Reparamétriser une chaîne d'ordre τ en chaîne d'ordre un implique généralement de définir un nouvel espace d'états tenant compte de toutes les séquences possibles de τ états consécutifs, soit une pige avec remise. Ainsi, cette opération sur une chaîne d'ordre τ admettant K états implique habituellement un espace d'états de cardinalité K^τ.

Puisque, dans le cas qui nous intéresse, nous sommes simplement intéressés par les séquences de longueurs $\ell \in \{1, \dots, \tau\}$ passées dans un même état, cette structure permet de réduire considérablement la cardinalité de l'espace d'états. Dans un modèle *DDMS* admettant K régimes et une dépendance aux séquences allant jusqu'à τ pas consécutifs—ci-après notés HMM-DDMS(K, τ)—, l'espace d'états est par conséquent de cardinalité $K\tau$. Afin de ne pas confondre l'état de la chaîne *originale* ($\{C_t\}_{t\in\mathbb{N}^\star}$) avec celui de la chaîne *instrumentale*, nous définissons la série $\{S_t\}_{t\in\mathbb{N}^\star}$ telle que

$$S_t \stackrel{\text{def}}{=} (C_t - 1)\tau + D_t \in \mathcal{S} = \{1, \dots, K\tau\}. \tag{5.2.2}$$

La définition d'un HMM-DDMS(K, τ) est présentée à la spécification du modèle *DDMS* (p.97).

Remarque 5.2.2 (Généralisation). Le modèle qu'ils proposent admet de plus une dépendance entre les valeurs décalées des rendements, comme à la spécification généralisée (p.58) des *HMMs* présentée au chapitre 3. Nous choisissons cependant de nous concentrer dans cette section sur l'incorporation de la durée dans la chaîne de Markov puisque l'inclusion de valeurs décalées du processus observé a déjà été étudiée à la section précédente sur les modèles *observation-driven*.

Spécification 6 : *Duration-Dependent Markov Switching models*

Distribution initiale	$C_1 \sim p(c_1)$
Probabilité de transition	$C_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t-1} = c_{1:t-1}) \sim p(c_t \mid c_{t-1}, d_{t-1})$
Probabilité d'émission	$X_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t} = c_{1:t}) \sim p(x_t \mid c_t)$

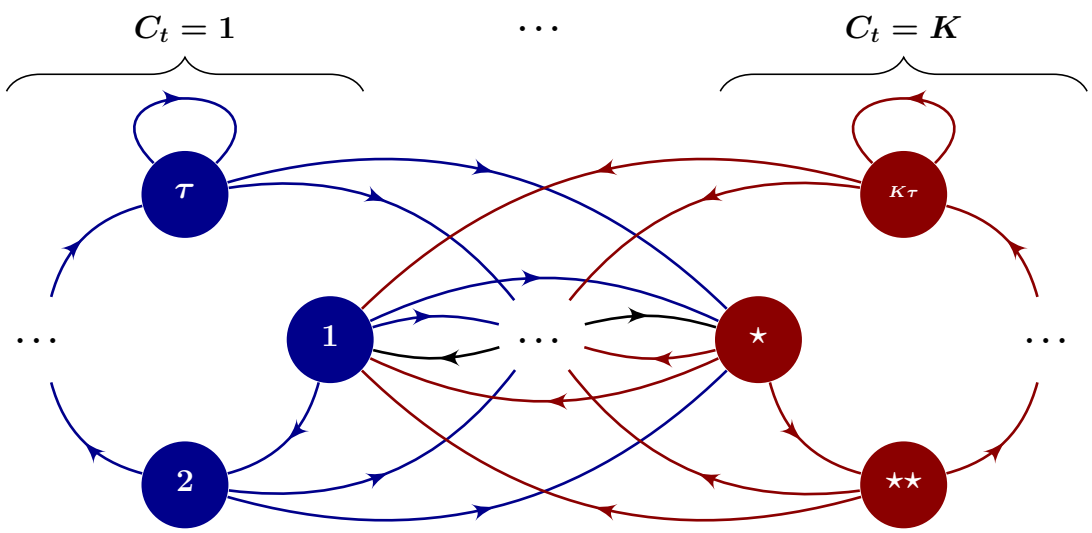
Nouvelle notation associée

$$D_t \overset{\text{def}}{=} \min\Big(\Big[D_{t-1} \cdot \mathbb{1}_{\{C_t = C_{t-1}\}} + 1\Big], \tau\Big), \qquad S_t \overset{\text{def}}{=} (C_t - 1)\tau + D_t \implies C_t = \left\lceil \tfrac{S_t}{\tau} \right\rceil,$$

$$\overset{\blacktriangledown}{Y}_{ij} \overset{\text{def}}{=} \mathbb{P}[S_{t+1} = j \mid S_t = i], \qquad \overset{\blacktriangledown}{\mathbf{u}}(t) \overset{\text{def}}{=} (\mathbb{P}[S_t = i])_{i \in \mathcal{S}},$$

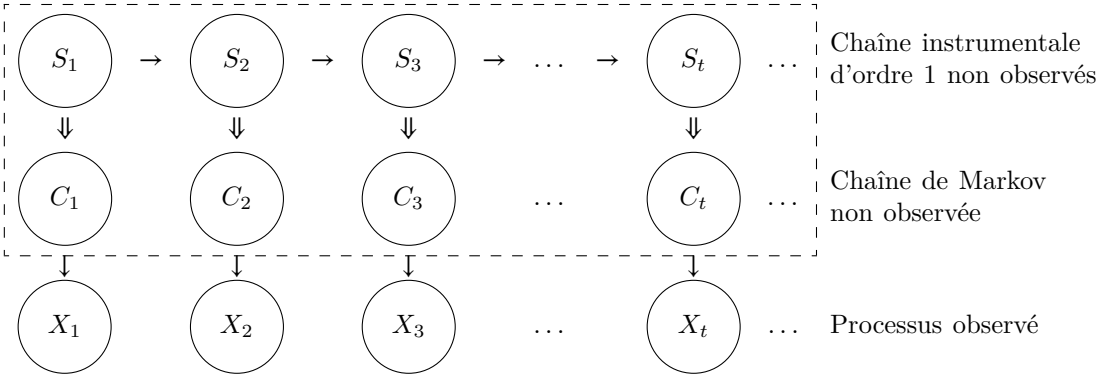
$$\overset{\blacktriangledown}{p}(x_t \mid x_{1:t-1}) \overset{\text{def}}{=} (\mathbb{P}[X_t = x_t \mid S_t = i, X_{1:t-1} = x_{1:t-1}])_{i \in \mathcal{S}}.$$

Représentation graphique du processus latent



$$\star = (K-1)\tau + 1,$$
$$\star\star = (K-1)\tau + 2.$$

Représentation graphique du modèle complet



5.2.1.1. Architecture de la matrice de transition

Les probabilités de transition de la chaîne *originale* sont obtenues, $\forall\, t \in \mathbb{N}^\star$, par les relations

$$\Upsilon_{ij}(t) \overset{\text{def}}{=} \mathbb{P}[\,C_{t+1} = j \mid C_t = i, D_t = d_t\,], \tag{5.2.3}$$

$$= \frac{\exp\left\{\beta_{ij}^0 + \beta_{ij}^1 \cdot d_t\right\}}{1 + \sum\limits_{\ell=1}^{K-1} \exp\{\beta_{i\ell}^0 + \beta_{i\ell}^1 \cdot d_t\}}, \quad \begin{cases} \forall\, i \in \{1, \ldots, K\} \text{ et} \\ \forall\, j \in \{1, \ldots, K-1\} \end{cases} \tag{5.2.4}$$

et

$$\Upsilon_{iK}(t) = 1 - \sum_{j=1}^{K-1} \Upsilon_{ij}(t), \tag{5.2.5}$$

$$= \frac{1}{1 + \sum\limits_{\ell=1}^{K-1} \exp\{\beta_{i\ell}^0 + \beta_{i\ell}^1 \cdot d_t\}}, \tag{5.2.6}$$

où l'utilisation de la fonction logistique permet à nouveau que les coefficients β_{ij}^0 et β_{ij}^1 soient non-contraints $\forall\, i, j \in \mathcal{C}$.

Les valeurs des coefficients β_{ij}^1 permettent de connaître le signe de la relation entre les probabilités de transition et la durée de temps dans un état. Lorsque la chaîne est par exemple dans l'état i,

$$\begin{aligned} \beta_{ii}^1 > 0 &\implies \text{l'état } i \text{ deviendra de plus en plus persistant,} \\ \beta_{ii}^1 < 0 &\implies \text{l'état } i \text{ deviendra de moins en moins persistant,} \\ \beta_{ii}^1 = 0 &\implies \text{la durée de temps dans l'état } i \text{ n'a pas d'effet} \\ &\qquad\quad \text{sur la probabilité de demeurer dans l'état } i. \end{aligned}$$

Remarque 5.2.3. Un modèle *DDMS* dont les coefficients β_{ij}^1 sont nuls $\forall i, j$ est équivalent à la spécification standard (p.56) d'un *HMM*.

Puisque c'est la chaîne de Markov associée à la variable S_t que nous devrons manipuler, nous redéfinissons les probabilités de transition afin de tenir compte de chaque valeur possible de la variable D_t. Ce faisant, la matrice de transition obtenue ne dépendra plus du temps t et sera donc homogène. Les probabilités de transition pour la chaîne *instrumentale* sont données par

$$\overset{\star}{\Upsilon}_{ij} \overset{\text{def}}{=} \mathbb{P}[\,S_{t+1} = j \mid S_t = i\,] = \mathbb{P}[\,C_{t+1} = j^\star, D_{t+1} = d_j \mid C_t = i^\star, D_t = d_i\,], \tag{5.2.7}$$

$$= \begin{cases} \frac{\exp\left\{\beta_{i^\star j^\star}^0 + \beta_{i^\star j^\star}^1 \cdot d_i\right\}}{1 + \sum\limits_{\ell=1}^{K-1} \exp\left\{\beta_{i^\star \ell}^0 + \beta_{i^\star \ell}^1 \cdot d_i\right\}}, & \begin{cases} \text{si } (i^\star = j^\star \text{ et } d_j = \min(d_i + 1, \tau)) \text{ ou} \\ \text{si } (i^\star \neq j^\star \text{ et } d_j = 1), \end{cases} \\ 0, \quad \text{sinon,} \end{cases} \tag{5.2.8}$$

avec $i^\star = \left\lceil \frac{i}{\tau} \right\rceil$, $j^\star = \left\lceil \frac{j}{\tau} \right\rceil$, $d_i = i - \tau(i^\star - 1)$ et $d_j = j - \tau(j^\star - 1)$. Ces probabilités de transition sont contenues dans la matrice

$$\overset{\blacksquare}{\mathbf{\Gamma}} \overset{\text{def}}{=} \left(\overset{\blacksquare}{\gamma}_{ij} \right)_{i,j \in \mathcal{S}}, \tag{5.2.9}$$

qui est construite par bloc.

Un des avantages de cette spécification réside dans le fait que, dans l'équation (5.2.8), les coefficients $\left(\beta^0_{i^\star j^\star}, \beta^1_{i^\star j^\star}\right)$, $\forall\, i^\star, j^\star \in \mathcal{C}$ sont définis en fonction de la chaîne *originale*. En effet, ceci implique que, puisque la chaîne *originale* admettait K régimes et que par conséquent sa matrice de transition avait $K(K-1)$ probabilités libres, le modèle requiert au total $(2K(K-1))$ paramètres pour peupler son processus de changement de régime, et ce peu importe la valeur de τ.

Puisque les probabilités de transition sont non-nulles seulement si l'état j est accessible depuis l'état i dans la chaîne *instrumentale*, il est clair en nous fiant à la représentation graphique du processus latent de la spécification du modèle *DDMS* (p.97) que la majorité d'entre elles seront nulles. À partir d'un état donné de la chaîne *instrumentale*, il y a deux cas possibles de transition :

(1) Augmentation du nombre de pas consécutifs dans le même état (jusqu'à un maximum de τ),
(2) Passage à un autre des $(K-1)$ régimes de la chaîne initiale et retour en $D_t = 1$.

La matrice $\overset{\blacksquare}{\mathbf{\Gamma}}$ est ainsi constituée de blocs correspondants au cas (1) sur sa diagonale, et de blocs correspondants au cas (2) autour. La forme que prennent les blocs sur la diagonale est présentée à la figure 5.1 (p.100) et celle pour les autres blocs est présentée à la figure 5.2 (p.101).

Exemple 5.2.1 (Construction de la matrice de transition pour un HMM-DDMS$(4,3)$)**.** *Soit un modèle* DDMS *bornant la dépendance à la durée en $\tau = 3$ et dans lequel la chaîne de Markov admet $K = 4$ régimes. La forme que prend alors la matrice de transition est présentée à la figure 5.3 (p.102).*

5.2.1.2. Adaptation de l'algorithme 1 : Filtre d'Hamilton (p.68)

Afin de pouvoir utiliser le filtre d'Hamilton pour les modèles *DDMSs*, nous devons redéfinir la distribution initiale $\vec{\mathfrak{u}}(1) \overset{\text{def}}{=} (\mathbb{P}[C_t = i])_{i \in \mathcal{C}}$ et la densité conditionnelle $\vec{p}(\, x_t \mid x_{1:t-1}) \overset{\text{def}}{=} (\mathbb{P}[X_t = x_t \mid X_{1:t-1} = x_{1:t-1}, C_t = i])_{i \in \mathcal{C}}$ pour qu'elles correspondent à la chaîne *instrumentale* S_t. Ainsi, en sachant que

$$\mathbb{P}[D_1 = d] = \begin{cases} 1, & \text{si } d = 1 \\ 0, & \text{sinon} \end{cases} \tag{5.2.10}$$

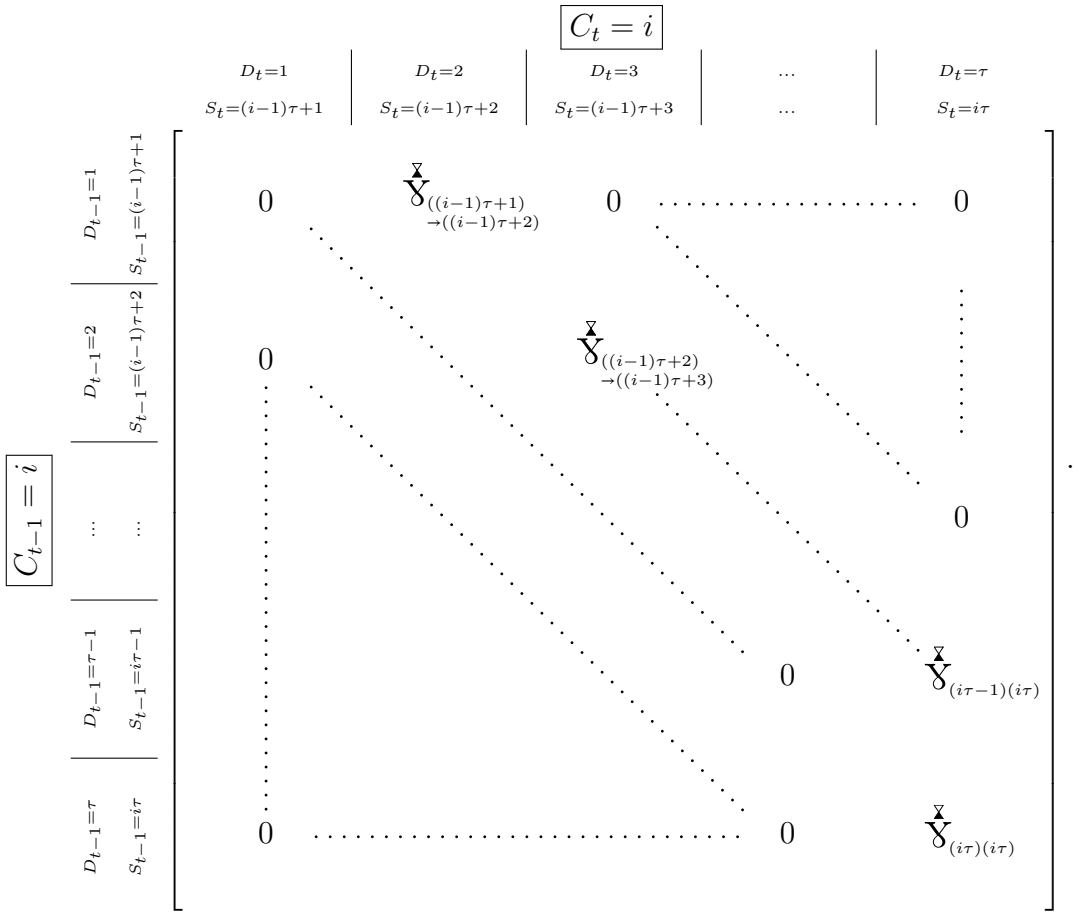


Figure 5.1 – Blocs sur la diagonale de $\overset{\blacktriangle}{\boldsymbol{\Gamma}}$ pour un HMM-DDMS(K, τ)

et donc que

$$\mathbb{P}[C_1 = i, D_1 = d_i] \overset{\text{Bayes}}{=\!=} \mathbb{P}[C_1 = i \mid D_1 = d_i]\mathbb{P}[D_1 = d_i], \tag{5.2.11}$$

$$\overset{\text{Ind.}}{=\!=} \mathbb{P}[C_1 = i]\mathbb{P}[D_1 = d_i], \tag{5.2.12}$$

$$= \begin{cases} \mathbb{P}[C_1 = i], & \text{si } d_i = 1 \\ 0, & \text{sinon,} \end{cases} \tag{5.2.13}$$

nous définissons dans un premier temps

$$\overset{\blacktriangle}{\boldsymbol{\mathfrak{u}}}(1) \overset{\text{def}}{=} (\mathbb{P}[S_1 = 1], \ldots, \mathbb{P}[S_1 = K\tau]), \tag{5.2.14}$$

$$= \Big(\mathbb{P}[C_1 = 1, D_1 = 1], \mathbb{P}[C_1 = 1, D_1 = 2], \ldots, \mathbb{P}[C_1 = 1, D_1 = \tau], \ldots,$$
$$\mathbb{P}[C_1 = K, D_1 = 1], \mathbb{P}[C_1 = K, D_1 = 2], \ldots, \mathbb{P}[C_1 = K, D_1 = \tau]\Big), \tag{5.2.15}$$

$$= \Big(\mathbb{P}[C_1 = 1, D_1 = 1], \overbrace{0, \ldots, 0}^{(\tau-1) \text{ fois}}, \ldots, \mathbb{P}[C_1 = K, D_1 = 1], \overbrace{0, \ldots, 0}^{(\tau-1) \text{ fois}}\Big), \tag{5.2.16}$$

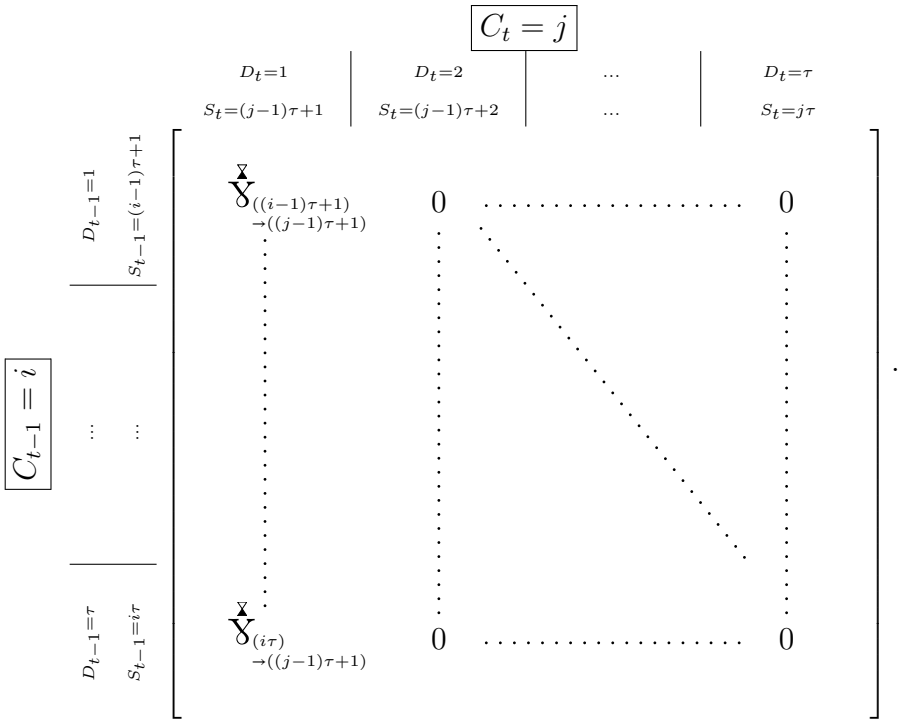


Figure 5.2 – Blocs hors de la diagonale de $\overset{\blacktriangledown}{\mathbf{\Gamma}}$ pour un HMM-DDMS(K, τ)

$$\overset{\text{Ind.}}{=} \Big(\mathbb{P}[C_1 = 1], \overbrace{0, \ldots, 0}^{(\tau-1) \text{ fois}}, \ldots, \mathbb{P}[C_1 = K], \overbrace{0, \ldots, 0}^{(\tau-1) \text{ fois}}\Big), \tag{5.2.17}$$

$$= \vec{\mathbf{u}}(1) \otimes \vec{e}_1, \tag{5.2.18}$$

avec $\vec{e}_1 \overset{\text{def}}{=} \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$, le premier vecteur de la base canonique de l'espace $\mathbb{R}^\tau$ et $\otimes$, le produit de Kronecker.

Pour ce qui est maintenant de remplacer la distribution conditionelle $\vec{p}(x_t \mid x_{1:t-1})$, puisque par hypothèse X_t est indépendant de D_t sachant C_t (i.e. $\mathbb{P}[X_t \mid C_t, D_t] \overset{\text{Hyp.}}{=\!=} \mathbb{P}[X_t \mid C_t]$), nous définissons

$$\overset{\blacktriangledown}{\vec{p}}(x_t \mid x_{1:t-1}) \overset{\text{def}}{=} (\mathbb{P}[X_t \mid S_t = 1, X_{1:t-1}], \ldots, \mathbb{P}[X_t \mid S_t = K\tau, X_{1:t-1}]), \tag{5.2.19}$$

$$\begin{aligned} &= \Big(\mathbb{P}[X_t \mid C_t = 1, D_t = 1, X_{1:t-1}], \ldots, \mathbb{P}[X_t \mid C_t = 1, D_t = \tau, X_{1:t-1}], \ldots, \\ &\qquad \mathbb{P}[X_t \mid C_t = K, D_t = 1, X_{1:t-1}], \ldots, \mathbb{P}[X_t \mid C_t = K, D_t = \tau, X_{1:t-1}]\Big), \\ &\overset{\text{Ind.}}{=} \Big(\overbrace{\mathbb{P}[X_t \mid C_t = 1, X_{1:t-1}], \ldots, \mathbb{P}[X_t \mid C_t = 1, X_{1:t-1}]}^{\tau \text{ fois}}, \ldots, \\ &\qquad \overbrace{\mathbb{P}[X_t \mid C_t = K, X_{1:t-1}], \ldots, \mathbb{P}[X_t \mid C_t = K, X_{1:t-1}]}^{\tau \text{ fois}}\Big), \end{aligned} \tag{5.2.20}$$

$$= \vec{p}(x_t \mid x_{1:t-1}) \otimes \mathbf{1}_\tau{}'. \tag{5.2.21}$$

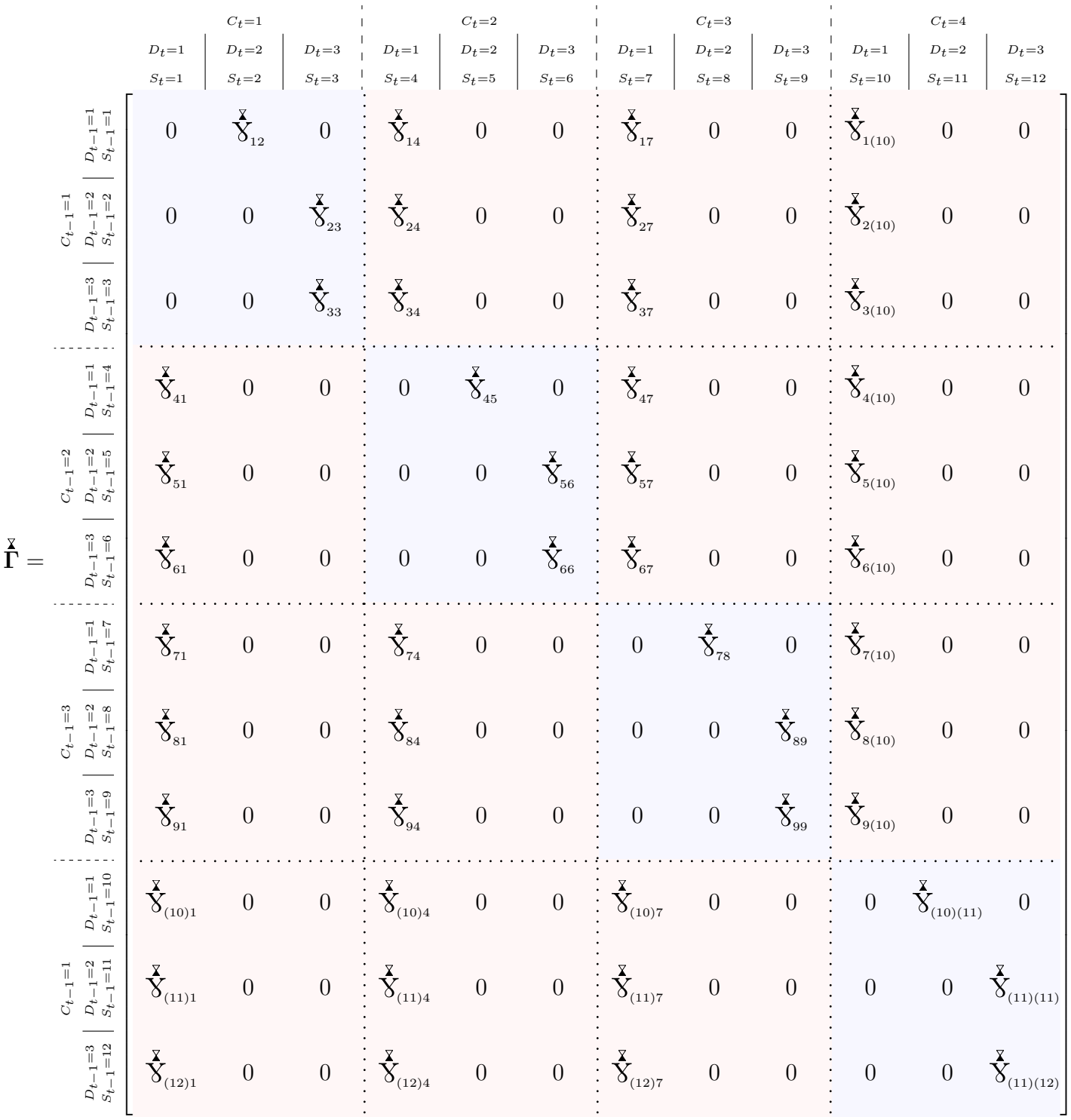


Figure 5.3 – Matrice $\mathbf{\Gamma}$ sous un HMM-DDMS($K = 4, \tau = 3$)

La version adaptée de cet algorithme tenant compte de ces nouvelles définitions est présentée à l'algorithme 2 : Filtre d'Hamilton pour modèle *DDMS* (p.103).

Algorithme 2 : Filtre d'Hamilton pour modèle *DDMS*

Soit la chaîne de Markov $\{S_t\}_{t\in\mathbb{N}^\star}$ définie en fonction de la chaîne de Markov $\{C_t\}_{t\in\mathbb{N}^\star}$ telle que

$$S_t \overset{\text{def}}{=} (C_t - 1)\tau + D_t \in \mathcal{S} = \{1, \ldots, K\tau\},$$

avec

$$D_t \overset{\text{def}}{=} \min\Big(D_{t-1}\mathbb{1}_{\{C_t = C_{t-1}\}} + 1, \tau\Big) \in \{1, \ldots, \tau\}.$$

Soit les notations suivantes associées à cette chaîne de Markov, ainsi que leurs interprétations probabilistes :

$$\vec{\overset{\blacktriangledown}{\mathbf{u}}}(1) \overset{\text{def}}{=} (\mathbb{P}[S_1 = 1], \ldots, \mathbb{P}[S_1 = K\tau]),$$

$$\vec{\overset{\blacktriangledown}{p}}(x_t \mid x_{1:t-1}) \overset{\text{def}}{=} (\mathbb{P}[X_t = x_t \mid S_t = 1, X_{1:t-1}], \ldots, \mathbb{P}[X_t = x_t \mid S_t = K\tau, X_{1:t-1}]),$$

$$\overset{\blacktriangledown}{\gamma}_{ij} \overset{\text{def}}{=} \mathbb{P}[S_{t+1} = j \mid S_t = i],$$

$$\overset{\blacktriangledown}{\mathbf{\Gamma}} \overset{\text{def}}{=} \Big(\overset{\blacktriangledown}{\gamma}_{ij}\Big)_{i,j\in\mathcal{S}},$$

et

$$\vec{a}_{S,t} \overset{\text{def}}{=} (\mathbb{P}[X_t, S_t = 1 \mid X_{1:t-1}], \ldots, \mathbb{P}[X_t, S_t = K\tau \mid X_{1:t-1}]),$$

$$a_t \overset{\text{def}}{=} \mathbb{P}[X_t \mid X_{1:t-1}],$$

$$\vec{\omega}_{S,t} \overset{\text{def}}{=} (\mathbb{P}[S_t = 1 \mid X_{1:t}], \ldots, \mathbb{P}[S_t = K\tau \mid X_{1:t}]).$$

Étape 1 : $t=1$

$$\vec{a}_{S,1} = \vec{\overset{\blacktriangledown}{\mathbf{u}}}(1) \odot \vec{\overset{\blacktriangledown}{p}}(x_1),$$

$$a_1 = \vec{a}_{S,1}\mathbf{1}_K,$$

$$\vec{\omega}_{S,1} = {}^{\vec{a}_{S,1}}\!/_{a_1}.$$

Étape 2 : $2\leq t\leq T$

$$\vec{a}_{S,t} = (\vec{\omega}_{S,t-1}\vec{\overset{\blacktriangledown}{\mathbf{\Gamma}}}) \odot \vec{\overset{\blacktriangledown}{p}}(x_t \mid x_{1:t-1}),$$

$$a_t = \vec{a}_{S,t}\mathbf{1}_K,$$

$$\vec{\omega}_{S,t} = {}^{\vec{a}_{S,t}}\!/_{a_t}.$$

Étape 3

$$\ell_T(\boldsymbol{\theta}) \quad = \quad \sum_{t=1}^{T} \log(a_t)$$

Coût computationnel	Référence
$\mathrm{O}(T(K\tau)^2)$	Maheu et McCurdy [2000a](p. 111-112)

Remarque

Le nombre d'opérations nécessaires pour compléter cet algorithme peut être réduit en construisant les vecteurs et matrices de sorte à profiter du fait que la majorité des entrées y sont nulles. Cet algorithme est valide même pour un modèle de Markov à variables latentes non-homogène tel que défini à la spécification généralisée (p.58).

Chapitre 6

Application

Dans ce mémoire, notre objectif était d'illustrer l'apport d'une reformulation hiérarchique et de la relaxation de l'hypothèse d'homogénéité à la modélisation de séries financières avec les modèles de Markov à variables latentes. Dans les chapitres précédents, nous avons présenté plusieurs méthodes permettant d'évaluer la performance et l'utilité de modèles statistiques. Nous proposons de séparer ces méthodes en quatre catégories :

(1) Qualité de l'ajustement aux données (log-vraisemblance, *AIC*, *AICc* et *BIC*). [1]
(2) Qualité de la prévision un pas en avant (*MAFE*, *MSFE*, *MASFE* et *MSSFE*). [2]
(3) Capacité à reproduire les faits stylisés présentés à la section 1.1 (p.35).
(4) Interprétabilité des paramètres.

Dans ce chapitre, nous rassemblerons les deux premiers critères dans une même section et comparerons la qualité de l'ajustement et des prévisions de tous les modèles présentés dans cet ouvrage sur deux séries, soit le niveau de l'indice S&P 500 [3] et le taux de change Canada-États-Unis (CADUSD) [4]. Cette portion de notre analyse se trouve à la section 6.1 (p.107). Ensuite, nous étudierons à la section 6.2 (p.111) la faculté des modèles entraînés sur le S&P 500 à reproduire les faits stylisés par le biais de séries simulées de 2 000 000 d'observations. Pour chacune de ces séries simulées, nous présenterons d'abord l'autocorrélation des log-rendements $\mathbb{C}\text{orr}[r_t, r_{t+h}]$, l'autocorrélation des rendements centrées en valeur absolue $\mathbb{C}\text{orr}[|\epsilon_t|, |\epsilon_{t+h}|]$ et au carré $\mathbb{C}\text{orr}\left[\epsilon_t^2, \epsilon_{t+h}^2\right]$—en plus de la différence entre les deux pour étudier la présence de l'effet de Taylor. Nous présenterons ensuite les coefficients d'asymétrie et d'applatissement et finalement la corrélation entre les log-rendements centrés et les chocs positifs $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{\epsilon_t, 0\}]$ et négatifs $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{-\epsilon_t, 0\}]$ afin d'étudier la présence de l'effet de levier. Afin d'illustrer le fait stylisé (3), soit le regroupement des volatilités,

1. Voir tableau 5.1.
2. Voir tableau 5.2.
3. (MSFT) [2020a].
4. (MSFT) [2020b].

nous étudierons les probabilités filtrées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$ et lissées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_T]$ [5] obtenues lors de l'entraînement des modèles sur le S&P 500. Finalement, nous nous intéresserons, à la section 6.3 (p.123), à l'interprétation des paramètres obtenus pour les modèles présentés aux chapitres 4 (p.75) et 5 (p.87).

Pour notre analyse, nous utilisons les log-rendements quotidiens du S&P 500 entre le 3 janvier 2000 et le 24 décembre 2020 et les différences pondérées du taux de change CADUSD entre le 18 septembre 2003 et le 25 décembre 2020. Afin d'évaluer la qualité de l'ajustement de nos modèles, nous avons choisi de séparer les échantillons en deux parties, soit un échantillon d'entraînement et un de test, représentant respectivement 90% et 10% de l'échantillon complet. Nous avons d'abord entraîné sur ces données une distribution normale univariée (UNIVARIÉE($K = 1$)), ainsi que trois modèles de mélange de $K \in \{2, 3, 4\}$ lois normales (MÉLANGE(K)). Ces modèles nous ont permis dans un premier temps d'obtenir des points de départ pour l'optimisation des modèles plus complexes et seront de plus utilisés comme modèles de référence. Nous avons ensuite entraîné trois HMM(K) correspondant à la spécification standard (p.56) avec $K \in \{2, 3, 4\}$, qui nous ont eux aussi aidé à calibrer les modèles plus complexes. Dans un quatrième temps, nous avons entraîné un HHMM($N = 2, k_1 = 2, k_2 = 2$) et un HHMM-simplifié($N = 2, k_1 = 2, k_2 = 2$), puis un modèle HMM-LOGIT($K = 4$) utilisant la série à l'étude comme variable explicative et un utilisant les log-rendements de l'indice de production industrielle totale américaine (INDPRO) [6], un modèle HMM-GAS($K = 2$) et finalement trois modèles HMM-DDMS($K = 4, \tau$), avec $\tau \in \{3, 5, 14\}$. L'optimisation a été effectuée grâce à la fonction `optim` disponible dans la librairie `stats` du langage `R`, avec un seuil de tolérance de 10^{-10} et en utilisant la méthode quasi-Newton *BFGS*.

Remarque 6.0.1. Nous utiliserons, dans les tableaux et graphiques de ce chapitre, les notations HHMM($K = 4$) et HHMM-simplifié($K = 4$) afin de faire référence aux modèles HHMM($N = 2, k_1 = k_2 = 2$) et HHMM-simplifié($N = 2, k_1 = k_2 = 2$) de manière plus concise. Ceci est simplement dû au fait que l'espace était une préoccupation puisque les tableaux étaient déjà de grandes tailles, et que les modèles hiérarchiques présentés admettaient tous deux régimes parents ($N = 2$) ayant chacun deux régimes enfants ($k_i = 2, \forall i$).

Remarque 6.0.2. L'estimation des paramètres du modèle HMM-GAS($K = 2$) a été possible grâce à la version 2.0 de la librairie `HMMGAS` rendu disponible par Marco Bazzi [2017]. Il a été nécessaire de modifier légèrement le code source de la librairie afin qu'elle permette que les volatilités puissent être distinctes dans chaque état. Cette librairie ne permet malheureusement que de considérer des modèles ayant $K = 2$ régimes, et ne permet pas de simuler des séquences en utilisant les paramètres estimés d'un modèle *GAS*. Ainsi, nous ne pourrons pas étudier leur capacité à reproduire les faits stylisés—outre le regroupement des volatilités.

5. Présentées dans l'appendice externe.
6. Federal Reserve System (US) [2020].

Remarque 6.0.3 (Accessibilité du code source). Le code `R` permettant de produire tous les résultats présentés dans cette section est disponible sur mon compte Github à l'adresse suivante : `https://github.com/GabrielLemyre/Code_R_Memoire_Maitrise`.

6.1. Qualité de l'ajustement et erreurs moyennes de prévision

Les mesures de qualité de l'ajustement basées sur la log-vraisemblance et l'erreur de prévision sur l'indice du S&P 500 sont présentées au tableau 6.1 (p.109) et ceux pour le taux de change Canada-États-Unis (CADUSD) sont présentées au tableau 6.2 (p.110).

Nous observons tout d'abord, en analysant les modèles n'admettant pas de dépendance temporelle et les *HMMs* standards, que nos résultats ne semblent pas a priori indiquer sans équivoque la présence de quatre régimes distincts, autant sur le S&P 500 que sur le CADUSD. L'apport d'un quatrième régime dans une structure standard semble en fait au mieux marginal. Une analyse des résultats obtenus sur les modèles plus complexes semblent cependant indiquer que quatre régimes sont observables dans la série. Ceci sera plus clair dans les sections subséquentes où nous nous intéresserons aux faits stylisés et à l'interprétation des paramètres.

Bien qu'ils ne soient en fait qu'une reformulation des *HMMs* standards, il semble que les modèles hiérarchiques se démarquent pour les mesures de qualité d'ajustement basées sur la log-vraisemblance, à la fois dans l'échantillon et à l'extérieur de celui-ci, et ce pour le S&P 500 et le CADUSD—à l'exception du *BIC* sur ce dernier. Une explication possible est que ces modèles sont potentiellement moins sensibles à une mauvaise initialisation puisque plusieurs paramètres entrent en jeu dans le calcul des probabilités de transition, rendant ces modèles généralement plus faciles à calibrer. Nous observons de plus que, puisqu'il ne requiert que 6 paramètres pour construire la matrice de transition contre 12 pour la version générale, les HHMM-simplifié($K = 4$) performent habituellement légèrement mieux que les modèles HHMM($K = 4$) lorsque le nombre de paramètres et la taille de l'échantillon sont des facteurs. La performance de ces modèles en terme d'erreur moyenne de prévision ressemble de près à la majorité des autres modèles étudiés sur le S&P 500, mais ils se démarquent pour l'erreur moyenne standardisée sur le CADUSD.

Puisque la performance des modèles HHMM($K = 4$) est comparable à celle des modèles HHMM-simplifié($K = 4$), mais que ces derniers nécessitent un peu moins de la moitié du temps d'entraînement sur le S&P 500 et un peu moins du tiers sur le CADUSD, ceux-ci sont probablement préférables comme premiers modèles afin d'évaluer l'apport potentiel d'une structure hiérarchique.

Pour les modèles HMM-LOGIT($K = 4$), il semble que la série INDPRO soit préférable aux valeurs décalées du S&P 500 lorsque la sélection de modèles se fait en terme de *BIC*, de *MAFE* ou de *MSFE*. Cependant, ce sont pourtant ces derniers qui dominent tous les autres modèles en terme d'erreur moyenne de prévision standardisée, et ce par une marge importante à l'intérieur et à l'extérieur de l'échantillon d'entraînement. L'inverse semble vrai sur la série du CADUSD, où c'est plutôt l'utilisation de valeurs décalées qui supplante l'utilisation de INDPRO. En effet, autant les mesures d'ajustement liées à la log-vraisemblance que celles liées à l'erreur moyenne de prévision standardisée indiquent que l'utilisation de valeurs décalées augmente significativement le pouvoir prédictif du modèle sur le CADUSD, ce qui est cohérent avec les résultats obtenus par Liu, Margaritis et Wang [2012].

Malgré le fait que les modèles HMM-GAS($K = 2$) ne soient grands vainqueurs dans aucune catégorie, leur performance en terme de *MAFE* et de *MSFE* est très semblable à la moyenne des autres modèles, et ils arrivent deuxième en terme de *MASFE* et de *MSSFE* sur le S&P 500. Il est de plus intéressant de noter que leur performance sur cette dernière série est supérieure hors de l'échantillon, creusant l'écart avec le modèle HMM-LOGIT($K = 4$) utilisant les valeurs décalées du S&P 500. Il se peut que ceci soit lié à l'observation de Bazzi, Blasques, Koopman et Lucas [2017], selon lesquels ces modèles ont une très grande faculté d'adaptation, et performent parfois mieux que le modèle ayant généré les observations. Sur le CADUSD, ces modèles performent relativement bien en terme de *BIC* et sensiblement comme les autres en terme de *MAFE* et de *MSFE*, mais leurs *MASFE* et *MSSFE* sont bien loin derrière les autres.

Finalement, les résultats obtenus pour les modèles HMM-DDMS($K = 4, \tau \in \{3, 5, 14\}$) ne semblent pas supporter l'idée que l'inclusion d'une dépendance à la durée permette à ces modèles de se démarquer en terme de log-vraisemblance ou d'erreur moyenne de prévision. En effet, ils se retrouvent dans la moyenne pour la majorité des mesures présentées dans ces tableaux, mais nécessitent généralement beaucoup plus de temps pour converger et l'apport d'une plus longue mémoire n'est pas évident.

En général, il semble que les modèles étudiés soient plus aptes à faire des prévisions sur l'indice du S&P 500 que sur le CADUSD. Ceci est potentiellement dû au fait que le CADUSD affiche des écarts à la moyenne moins importants, ne justifiant donc pas forcément l'utilisation de modèles si complexes.

Résultats **DANS l'échantillon 2000-01-03 au 2018-11-20**

	$\ell_T(\widehat{\boldsymbol{\theta}})$	**AIC**	**AICc**	**BIC**	**MAFE**	**MSFE**	**MASFE**	**MSSFE**	D_K	**Temps**
UNIVARIÉE($K=1$)	−7625.2	15254.5	15254.5	15267.4	0.7998	1.4488	0.5520	0.6901	2	178ms
MÉLANGE($K=2$)	−7023.9	14057.9	14057.9	14090.2	0.7998	1.4488	0.5520	0.6901	5	11s
MÉLANGE($K=3$)	−6952.7	13921.3	13921.4	13973.1	0.7998	1.4488	0.5520	0.6901	8	51s
MÉLANGE($K=4$)	−6942.4	13906.8	13906.9	13978.0	0.7998	1.4488	0.5520	0.6901	11	2m 19s
HMM-Homogène($K=2$)	−6665.7	13343.3	13343.3	13382.1	0.7998	1.4488	0.5527	0.6925	6	14s
HMM-Homogène($K=3$)	−6453.6	12931.3	12931.3	13008.9	0.7998	1.4487	0.5508	0.6870	12	1m 56s
HMM-Homogène($K=4$)	−6463.4	12966.9	12967.1	13096.2	0.7998	1.4487	0.5536	0.6945	20	2m 55s
HHMM($K=4$)	−6401.3	12842.5	12842.7	12971.9	0.7998	1.4487	0.5452	0.6749	20	5m 6s
HHMM-simplifié($K=4$)	−6418.6	12865.2	12865.3	12955.7	0.7997	1.4487	0.5517	0.6911	14	1m 57s
LOGIT($K=4$) ∼ **S&P 500**[−1]	−6470.7	13005.445	13005.9	13212.4	0.8219	1.4740	0.1230	0.0390	32	16m 52s
LOGIT($K=4$) ∼ **INDPRO**	−6427.3	12918.5	12919.0	13125.4	0.7993	1.4461	0.6040	0.7347	32	18m 52s
GAS($K=2$)	−6597.4	13214.7	13214.8	13279.4	0.8002	1.4522	0.3552	0.4353	10	5s
DDMS($K=4,\tau=3$)	−6447.5	12959.0	12959.5	13166.0	0.7998	1.4489	0.5500	0.6851	32	18m 3s
DDMS($K=4,\tau=5$)	−6484.1	13032.3	13032.7	13239.2	0.8000	1.4500	0.5513	0.6885	32	33m 11s
DDMS($K=4,\tau=14$)	−6444.9	12953.8	12954.2	13160.7	0.7999	1.4488	0.5357	0.6527	32	52m 39s

Résultats **HORS de l'échantillon 2018-11-21 au 2020-12-24**

	$\ell_T(\widehat{\boldsymbol{\theta}})$	**AIC**	**AICc**	**BIC**	**MAFE**	**MSFE**	**MASFE**	**MSSFE**	D_K	**Temps**
UNIVARIÉE($K=1$)	−1077.8	2159.6	2159.6	2168.1	0.9779	2.7261	0.6749	1.2985	2	178ms
MÉLANGE($K=2$)	−887.0	1784.0	1784.1	1805.3	0.9779	2.7261	0.6749	1.2985	5	11s
MÉLANGE($K=3$)	−871.2	1758.3	1758.6	1792.5	0.9779	2.7261	0.6749	1.2985	8	51s
MÉLANGE($K=4$)	−864.7	1751.4	1751.9	1798.3	0.9779	2.7261	0.6749	1.2985	11	2m 19s
HMM-Homogène($K=2$)	−840.4	1692.9	1693.1	1718.5	0.9779	2.7261	0.6750	1.2987	6	14s
HMM-Homogène($K=3$)	−787.8	1599.7	1600.3	1650.9	0.9779	2.7261	0.6744	1.2964	12	1m 56s
HMM-Homogène($K=4$)	−784.1	1608.1	1609.8	1693.5	0.9779	2.7261	0.6750	1.2985	20	2m 55s
HHMM($K=4$)	−767.5	1575.0	1576.7	1660.4	0.9779	2.7261	0.6771	1.3068	20	5m 6s
HHMM-simplifié($K=4$)	−773.8	1575.6	1576.5	1635.4	0.9779	2.7260	0.6866	1.3440	14	1m 57s
LOGIT($K=4$) ∼ **S&P 500**[−1]	−801.7	1667.4	1671.7	1804.0	1.0078	2.7625	0.1721	0.0748	32	16m 52s
LOGIT($K=4$) ∼ **INDPRO**	−788.9	1641.9	1646.1	1778.4	0.9778	2.7164	0.5056	0.5525	32	18m 52s
GAS($K=2$)	−837.0	1693.9	1694.3	1736.6	0.9821	2.7380	0.2656	0.2187	10	5s
DDMS($K=4,\tau=3$)	−791.1	1646.3	1650.6	1782.9	0.9780	2.7261	0.6747	1.2976	32	18m 3s
DDMS($K=4,\tau=5$)	−796.1	1656.3	1660.5	1792.8	0.9779	2.7261	0.6777	1.3092	32	33m 11s
DDMS($K=4,\tau=14$)	−789.3	1642.7	1646.9	1779.2	0.9783	2.7264	0.6517	1.2097	32	52m 39s

Tableau 6.1 – Résultats sur **S&P 500**

Résultats **DANS l'échantillon 2003-09-18 au 2019-04-08**

	$\ell_T\left(\widehat{\boldsymbol{\theta}}\right)$	**AIC**	**AICc**	**BIC**	**MAFE**	**MSFE**	**MASFE**	**MSSFE**	D_K	**Temps**
UNIVARIÉE($K=1$)	−3565.1	7134.1	7134.2	7146.8	0.4262	0.3412	1.2490	2.9305	2	181ms
MÉLANGE($K=2$)	−3381.9	6773.8	6773.9	6805.4	0.4262	0.3412	1.2490	2.9305	5	9s
MÉLANGE($K=3$)	−3371.2	6758.5	6758.5	6808.9	0.4262	0.3412	1.2490	2.9305	8	42s
MÉLANGE($K=4$)	−3370.0	6762.1	6762.1	6831.4	0.4262	0.3412	1.2490	2.9305	11	28s
HMM-Homogène($K=2$)	−3236.1	6484.3	6484.3	6522.1	0.4262	0.3412	1.2509	2.9387	6	24s
HMM-Homogène($K=3$)	−3176.8	6377.6	6377.7	6453.3	0.4262	0.3412	1.2658	3.0170	12	1m 40s
HMM-Homogène($K=4$)	−3155.9	6351.9	6352.1	6478.0	0.4262	0.3413	1.2698	3.0464	20	2m 54s
HHMM($K=4$)	−3153.9	6347.9	6348.1	6474.0	0.4262	0.3412	1.2558	2.9702	20	6m 40s
HHMM-simplifié($K=4$)	−3166.0	6360.1	6360.2	6448.3	0.4275	0.3422	0.7610	1.1181	14	2m 3s
LOGIT($K=4$) ∼ **CADUSD**[−1]	−3163.6	6391.3	6391.8	6593.0	0.4261	0.3411	0.8653	1.3313	32	20m 58s
LOGIT($K=4$) ∼ **INDPRO**	−3182.8	6429.6	6430.2	6631.4	0.4260	0.3412	1.2758	2.9571	32	12m 20s
GAS($K=2$)	−3198.7	6417.4	6417.4	6480.4	0.4263	0.3413	3.5914	29.4428	10	2s
DDMS($K=4,\tau=3$)	−3208.0	6480.1	6480.6	6681.9	0.4262	0.3412	1.2606	2.9926	32	6m 13s
DDMS($K=4,\tau=5$)	−3160.3	6384.6	6385.1	6586.3	0.4261	0.3411	1.2625	3.0033	32	24m 15s
DDMS($K=4,\tau=14$)	−3155.8	6375.6	6376.1	6577.4	0.4262	0.3412	1.2644	3.0065	32	34m 40s

Résultats **HORS de l'échantillon 2019-04-09 au 2020-12-25**

	$\ell_T\left(\widehat{\boldsymbol{\theta}}\right)$	**AIC**	**AICc**	**BIC**	**MAFE**	**MSFE**	**MASFE**	**MSSFE**	D_K	**Temps**
UNIVARIÉE($K=1$)	−290.3	584.5	584.5	592.7	0.3003	0.1819	0.8799	1.5622	2	181ms
MÉLANGE($K=2$)	−251.6	513.2	513.3	533.7	0.3003	0.1819	0.8799	1.5622	5	9s
MÉLANGE($K=3$)	−249.1	514.2	514.5	547.1	0.3003	0.1819	0.8799	1.5622	8	42s
MÉLANGE($K=4$)	−248.4	518.9	519.5	564.1	0.3003	0.1819	0.8799	1.5622	11	28s
HMM-Homogène($K=2$)	−228.7	469.3	469.5	494.0	0.3003	0.1819	0.8805	1.5642	6	24s
HMM-Homogène($K=3$)	−222.1	468.2	468.9	517.4	0.3003	0.1819	0.8796	1.5610	12	1m 40s
HMM-Homogène($K=4$)	−209.0	457.9	459.9	540.0	0.3003	0.1819	0.8790	1.5590	20	2m 54s
HHMM($K=4$)	−203.4	446.8	448.8	528.9	0.3003	0.1819	0.8921	1.6059	20	6m 40s
HHMM-simplifié($K=4$)	−213.7	455.4	456.3	512.8	0.3019	0.1838	0.5060	0.5161	14	2m 3s
LOGIT($K=4$) ∼ **CADUSD**[−1]	−224.0	512.1	517.2	643.4	0.3007	0.1825	0.6120	0.7127	32	20m 58s
LOGIT($K=4$) ∼ **INDPRO**	−285.5	635.0	640.0	766.3	0.3070	0.1905	1.0248	2.1868	32	12m 20s
GAS($K=2$)	−219.3	458.5	459.0	499.6	0.3003	0.1819	4.6040	47.2643	10	2s
DDMS($K=4,\tau=3$)	−221.7	507.4	512.5	638.8	0.3003	0.1819	0.8958	1.6191	32	6m 13s
DDMS($K=4,\tau=5$)	−222.8	509.5	514.6	640.9	0.3003	0.1819	0.8979	1.6266	32	24m 15s
DDMS($K=4,\tau=14$)	−218.9	501.9	507.0	633.2	0.3003	0.1819	0.8949	1.6159	32	34m 40s

Tableau 6.2 – Résultats sur **CADUSD**

6.2. Reproduction des faits stylisés avec les modèles entraînés sur le S&P 500

Rappelons d'abord que la forme donnée aux log-rendements du S&P 500 à la section 1.1 (p.35) était

$$r_t = \mu_t + \epsilon_t = \mu_t + \eta_t \sigma_t, \tag{6.2.1}$$

avec

$$\mu_t = \mathbb{E}[r_t \mid r_{1:t-1}], \tag{6.2.2}$$

$$\sigma_t^2 = \mathbb{V}ar[r_t \mid r_{1:t-1}], \tag{6.2.3}$$

$$\eta_t \sim \mathcal{N}(0,1) \tag{6.2.4}$$

et où les ϵ_t correspondent aux log-rendements centrées, ou à l'erreur statistique.

Rappelons de plus que les faits stylisés sont

(1) l'absence d'autocorrélation dans les rendements (i.e. $\mathbb{C}\text{orr}[r_t, r_{t+h}] \approx 0, \forall h \geq 1$),
(2) la présence d'autocorrélation positive dans les processus $\{\epsilon_t^2\}$ et $\{|\epsilon_t|\}$,
(3) le regroupement (*clustering*) des volatilités,
(4) des ailes épaisses et une asymétrie négative pour la distribution des log-rendements ($\tilde{\mu}_3 < 0$ et $\tilde{\mu}_4 > 3$),
(5) l'effet de levier : $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{-\epsilon_t, 0\}] > \mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{\epsilon_t, 0\}]$,

en plus de l'effet de Taylor observable en calculant $\mathbb{C}\text{orr}[|\epsilon_t|, |\epsilon_{t+h}|] - \mathbb{C}\text{orr}\left[\epsilon_t^2, \epsilon_{t+h}^2\right]$.

Tel que mentionné en introduction de chapitre, nous avons, afin d'illuster la faculté des divers modèles à reproduire ces faits stylisés, simulé pour chaque modèle entraîné sur le S&P 500 une série de 2 000 000 d'observations. Nous comparons ensuite les diverses quantités d'intérêt à celles obtenues empiriquement sur la série des log-rendements du S&P 500. Les sections qui suivent sont ordonnées afin de mouler la liste de faits stylisés présentée ci-haut.

6.2.1. Absence d'autocorrélation dans les rendements

Les fonctions d'autocorrélation $\mathbb{C}\text{orr}[r_t, r_{t+h}]$ pour $h \in \{1, \ldots, 200\}$ sont présentées à la figure 6.1 (p.112). Au premier regard, il semble que les séries simulées par tous les modèles affichent une absence d'autocorrélation dans les log-rendements. Ainsi, tous les modèles étudiés reproduisent aisément le fait stylisé (1) (p.36).

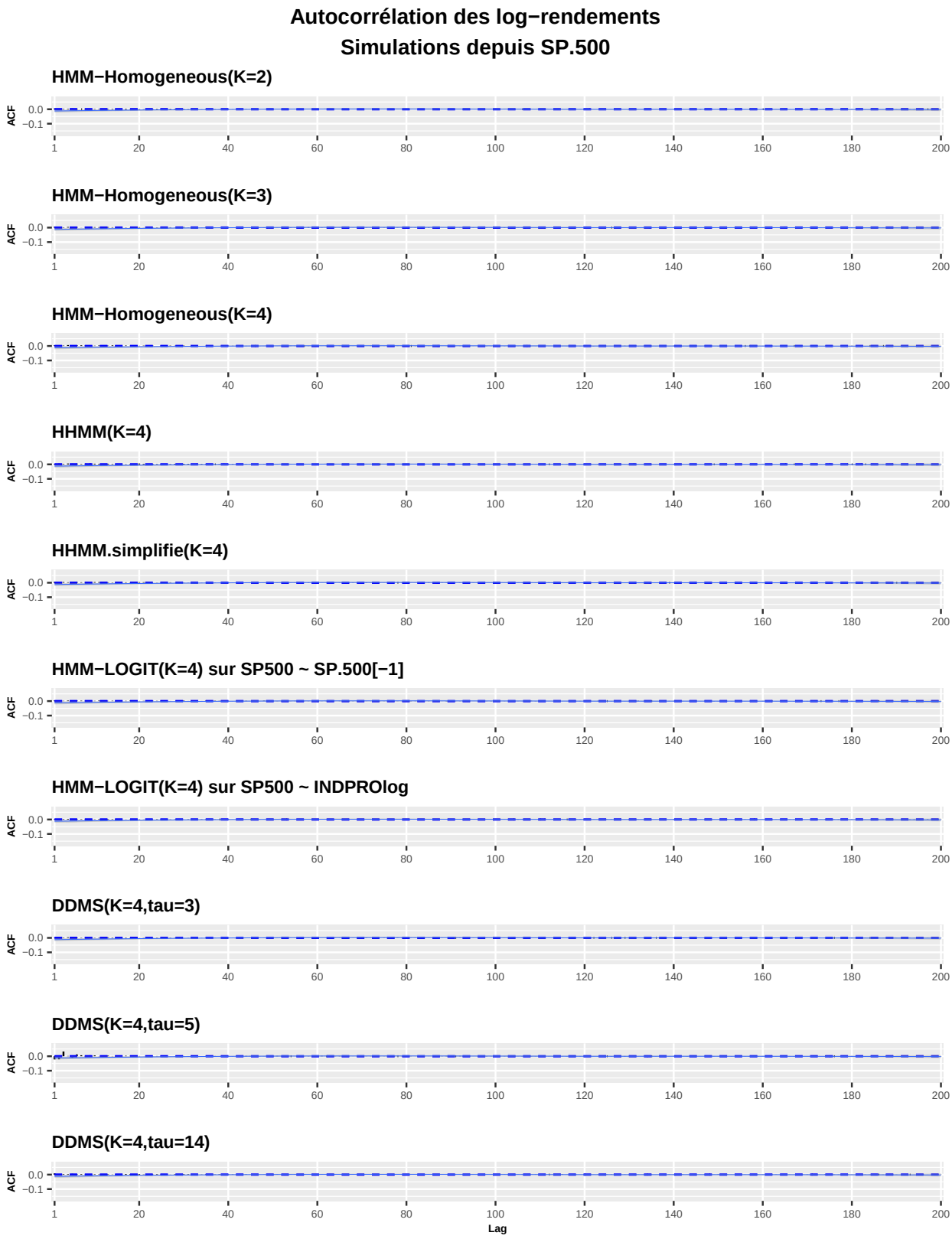


Figure 6.1 – Fonctions d'autocorrélation des log-rendements du S&P 500

6.2.2. Présence d'autocorrélation positive dans les processus $\{|\epsilon_t|\}$ et $\{\epsilon_t^2\}$

Les fonctions d'autocorrélation des log-rendements centrées en valeur absolue $\mathbb{C}\text{orr}[|\epsilon_t|, |\epsilon_{t+h}|]$ pour $h \in \{1, \ldots, 200\}$ sont présentées à la figure 6.2 (p.114) et celles pour les log-rendements centrées au carré $\mathbb{C}\text{orr}\left[\epsilon_t^2, \epsilon_{t+h}^2\right]$ sont présentées à la figure 6.3 (p.115). Nous avons superposé, sur chacun de ces graphiques, les fonctions d'autocorrélation empirique lissées correspondantes afin de pouvoir comparer celle-ci à celles des séries simulées. Bien que selon Rydén, Teräsvirta et Åsbrink [1998] ce fait stylisé soit le plus difficile à reproduire avec les modèles de Markov à variables latentes, il semble que la relaxation de l'hypothèse d'homogénéité ainsi qu'une architecture hiérarchique puisse permettre aux *HMMs* de reproduire celui-ci plus fidèlement.

En effet, parmi les modèles homogènes, les modèles hiérarchiques affichent généralement une plus grande persistance dans ces deux processus, avec le modèle HHMM($K = 4$) moulant presque exactement la fonction d'autocorrélation empirique lissée sur un grand horizon. Parmi les modèles non-homogènes, les HMM-DDMS($K = 4, \tau \in \{3, 14\}$) sont eux aussi très persistants, ce qui était attendu puisque ces derniers incorporent directement la persistance dans un régime comme variable explicative. De plus, bien que ce soit marginal, l'augmentation de la mémoire du modèle semble augmenter aussi la persistance dans ces deux processus. Ainsi, autant les modèles hiérarchiques que les *DDMSs* affichent une décroissance bien plus lente de ces fonctions d'autocorrélation que les *HMMs* standards.

Pour ce qui est de l'effet de Taylor, la différence entre les fonctions d'autocorrélation de $\{|\epsilon_t|\}$ et de $\{\epsilon_t^2\}$ est présentée à la figure 6.4 (p.116)—sur laquelle la fine ligne bleue représente cette différence sur la série du S&P 500. Il semble que tous les modèles permettent de simuler des séries respectant l'effet de Taylor puisque nous avons $\mathbb{C}\text{orr}\left[\epsilon_t^2, \epsilon_{t+h}^2\right] < \mathbb{C}\text{orr}[|\epsilon_t|, |\epsilon_{t+h}|]$, $\forall h > 0$, et ce pour toutes les séries. Cet effet est à nouveau beaucoup plus apparent pour les modèles hiérarchiques et les *DDMSs*, mais aussi pour le HMM-LOGIT($K = 4$) utilisant les valeurs décalées du S&P 500. Il est intéressant de noter ici que c'est le HHMM-simplifié($K = 4$) qui affiche la plus lente décroissance, moulant de très près l'effet de Taylor observé empiriquement.

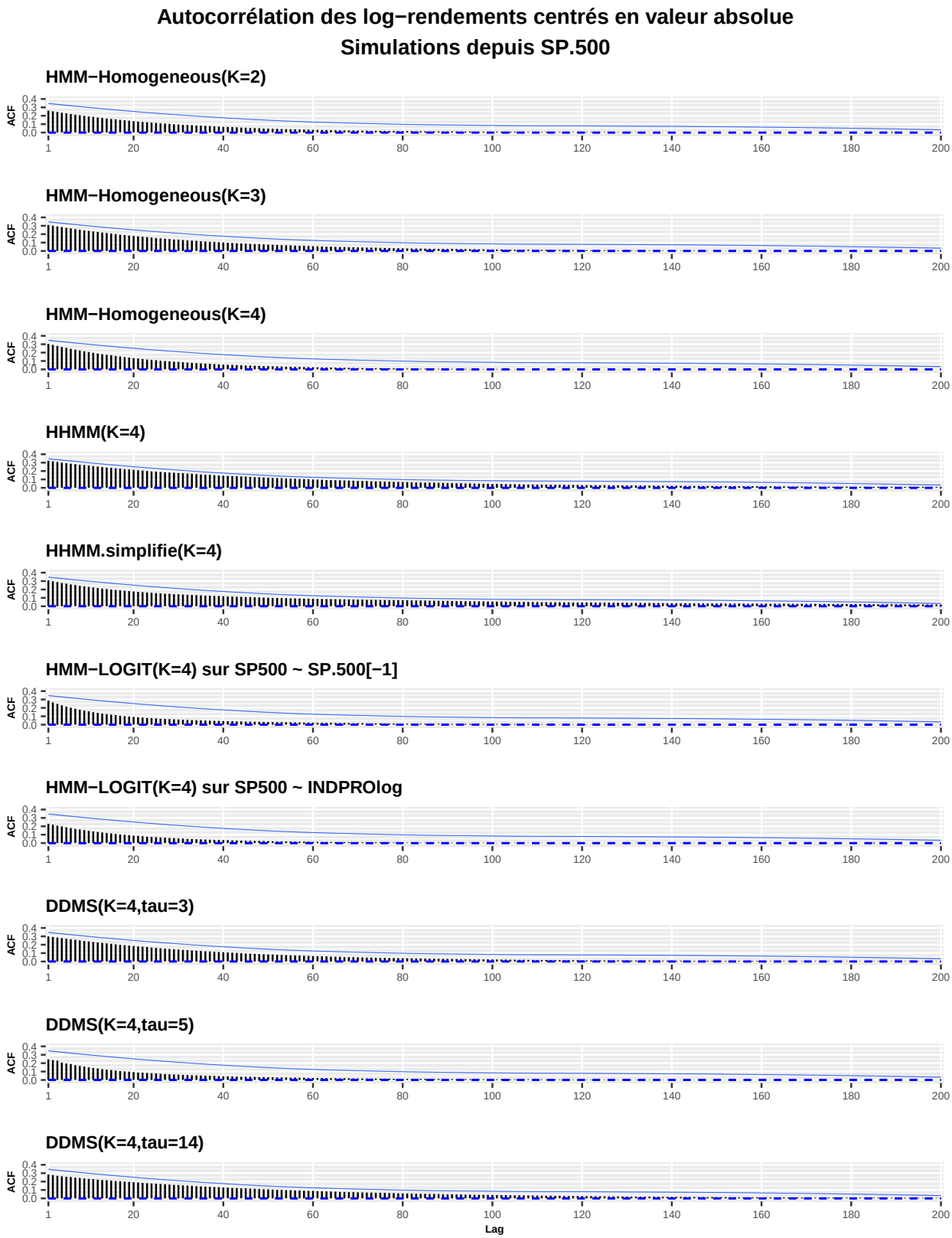


Figure 6.2 – Fonctions d'autocorrélation des log-rendements centrés en valeur absolue $\{|\epsilon_t|\}$ du S&P 500 et des séries simulées avec les modèles à l'étude

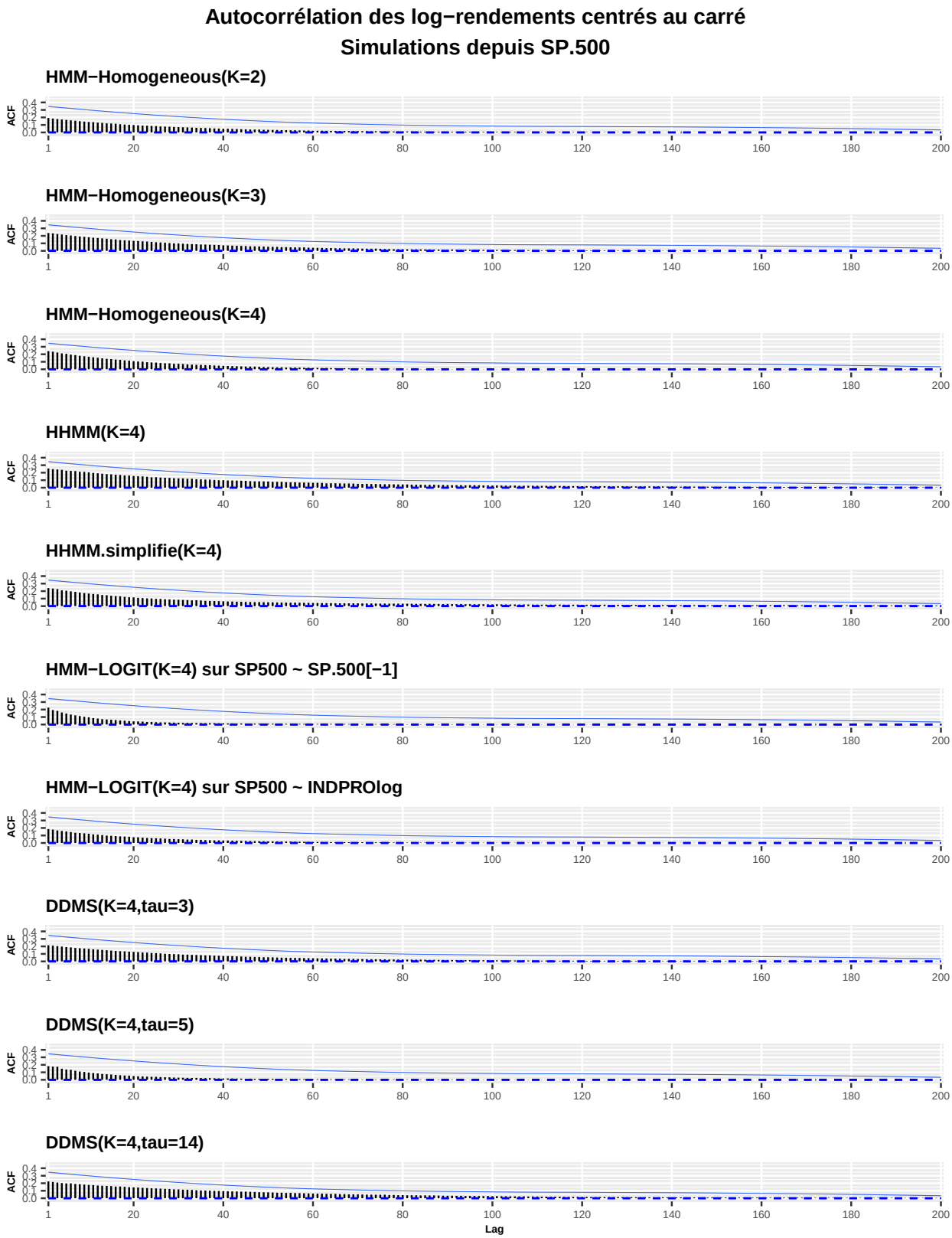


Figure 6.3 – Fonctions d'autocorrélation des log-rendements centrés et au carré $\{\epsilon_t^2\}$ du S&P 500 et des séries simulées avec les modèles à l'étude

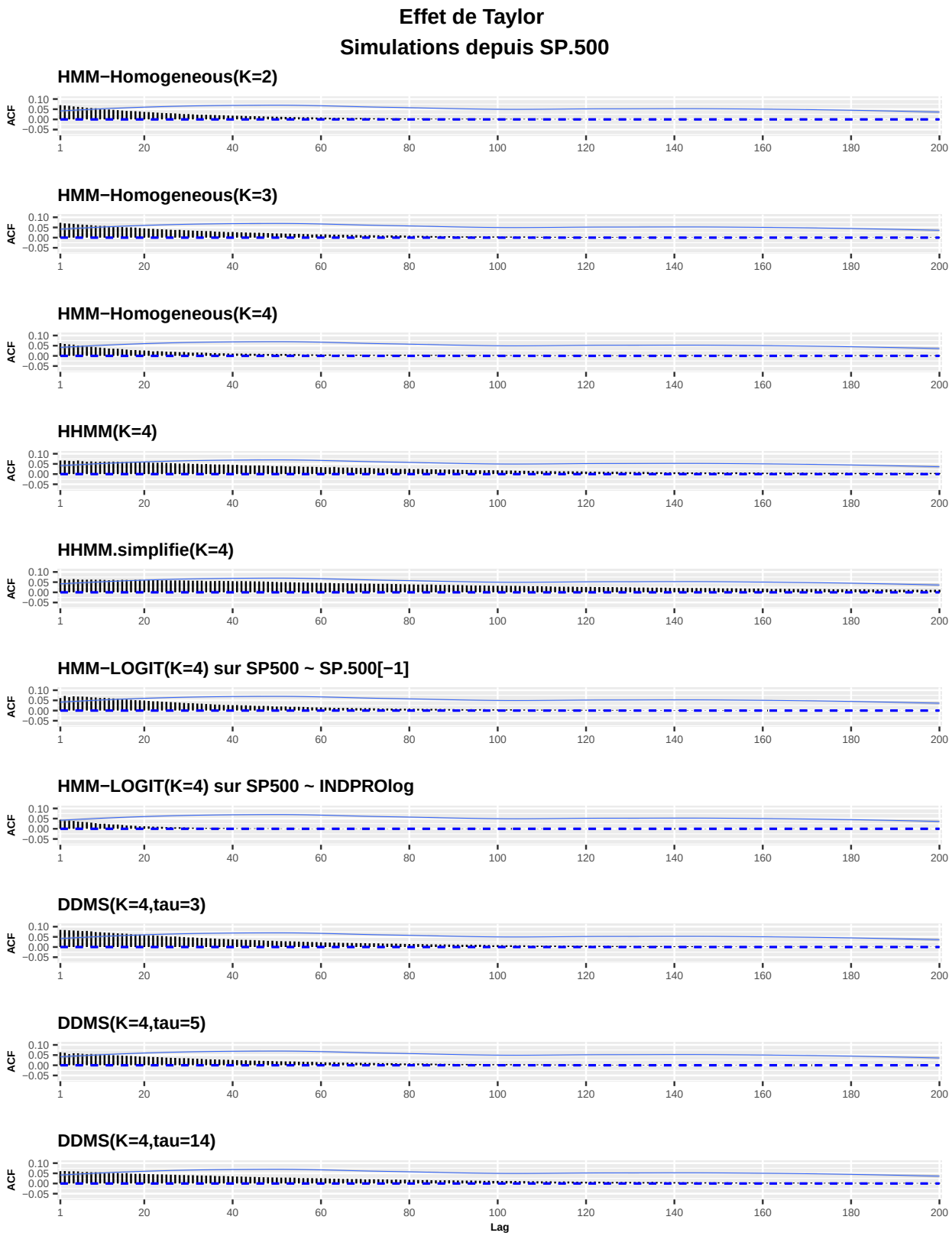


Figure 6.4 – Illustration de l'effet de Taylor sur le S&P 500 et les séries simulées avec les modèles à l'étude

6.2.3. Regroupement des volatilités

Afin d'illustrer le regroupement des volatilités, nous présentons à la figure 6.5 (p.118) les probabilités filtrées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$ obtenues par le filtre d'Hamilton et introduites à la section 3.6 (p.66), et à la figure 6.6 (p.119) les probabilités lissées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_T]$ obtenues par le filtre de Kim et présentées dans l'appendice externe. Ces graphiques présentent des régions correspondantes à un empilement vertical des probabilités filtrées ou lissées à chaque pas de temps et les lignes pointillées verticales indiquent la position de la dernière date de l'échantillon d'entraînement. Les états ont de plus été ordonnés de sorte que l'état le plus en bas (en rouge) soit toujours le plus volatil et ainsi de suite jusqu'au moins volatil tout au dessus (en bleu). Nous reportons de plus en annexe A.1 (p.140) des graphiques plus détaillés pour chaque modèle.

Bien que cet effet de regroupement soit apparent en regardant les probabilités filtrées, celui-ci est habituellement beaucoup plus évident en observant les probabilités lissées puisque les régions y sont plus définies. Ces graphiques semblent suggérer que tous les modèles étudiés permettent un regroupement des volatilités. En effet, les régions colorées sont habituellement assez larges, et ce même en période de haute volatilité. L'augmentation du nombre de régimes a pour effet de réduire la largeur de ces bandes puisque les régimes supplémentaires ont tendances à ajouter plus de granularité à l'étude de la volatilité. Ceci sera plus évident à la section 6.3 (p.123) où nous présenterons les valeurs estimées des paramètres.

Remarque 6.2.1. Nous remarquons dès lors en observant ces graphiques que les probabilités filtrées et lissées d'un régime des modèles *DDMSs* semblent ne jamais dépasser la barre du 50%. Ceci sera discuté à la section 6.3.6 (p.127) sur l'interprétation des paramètres.

Remarque 6.2.2. Nous observons de plus que bien que l'année 2020 ait été difficile, la bonne nouvelle est que tous les modèles semblent indiquer que nous entrons dans une période de plus faible turbulence sur les marchés boursiers, répit dont nous avons tous bien besoin.

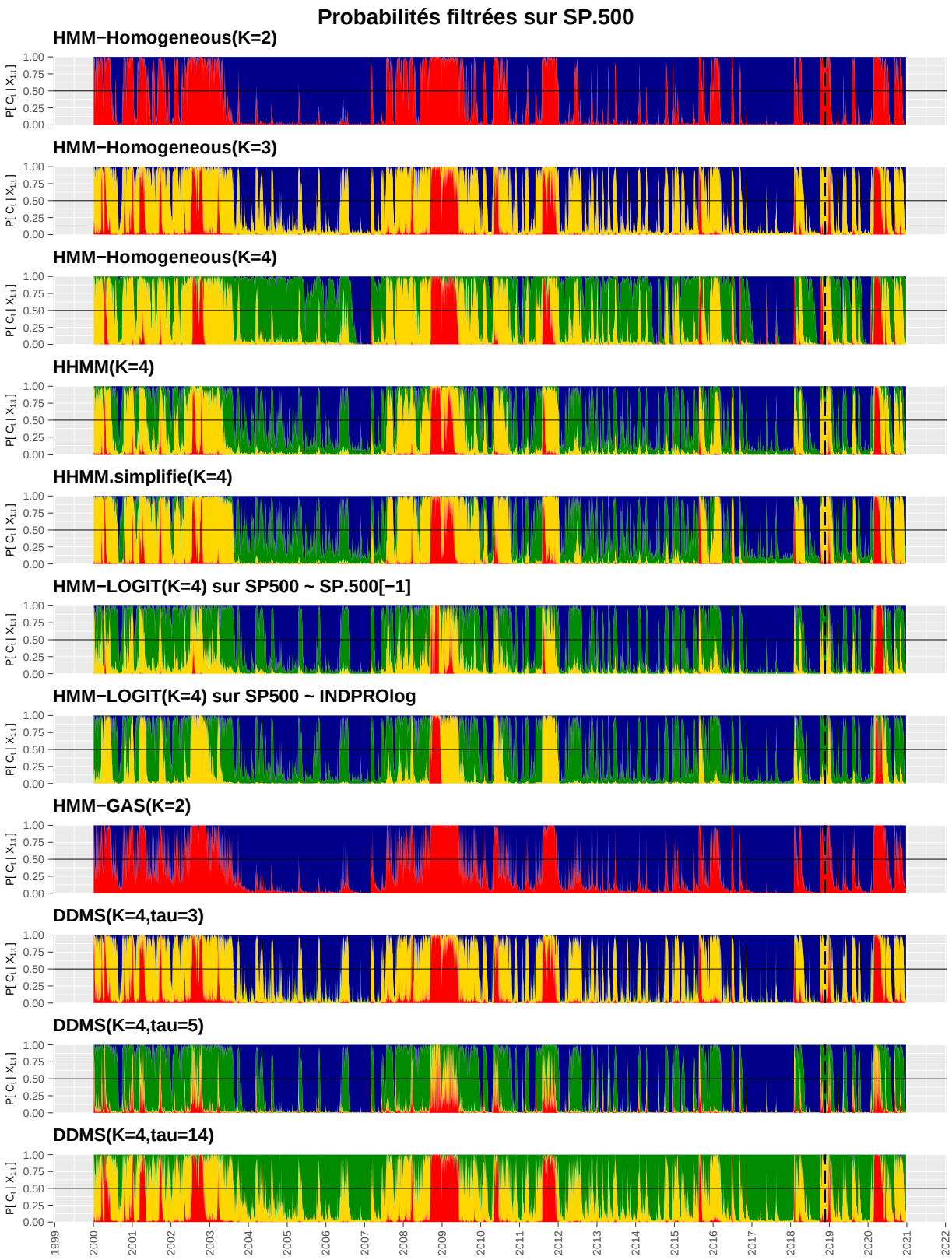


Figure 6.5 – Probabilités filtrées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$ estimées sur le S&P 500

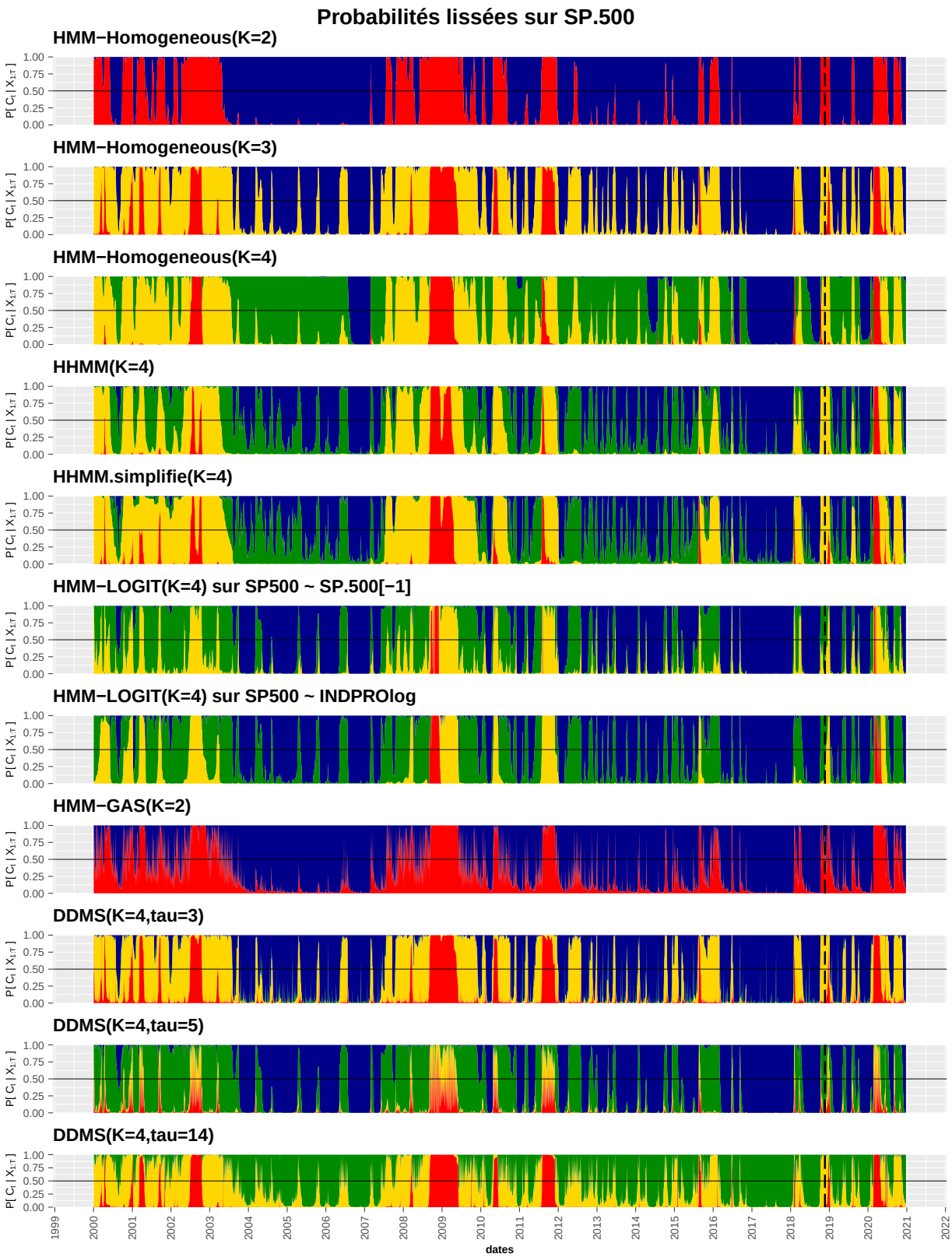


Figure 6.6 – Probabilités lissées $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_T]$ estimées sur le S&P 500

6.2.4. Ailes épaisses et asymétrie négative

Afin d'illustrer ce fait stylisé, nous présentons au tableau 6.3 les même statistiques descriptives que celles présentées au tableau 1.1 de la section 1.1 (p.35). Tous les modèles étudiés affichent un coefficient d'asymétrie négatif d'une ampleur comparable au coefficient empirique, avec le coefficient $\tilde{\mu}_3$ du HHMM-simplifié($K = 4$) étant le plus près. Il en va de même pour le coefficient d'aplatissement $\tilde{\mu}_4$ de nos modèles qui sont tous supérieurs à 3, avec cette fois le HHMM($K = 4$) en première place. Il semble donc que tous les modèles étudiés sachent reproduire ce fait stylisé relativement bien.

Tableau 6.3 – Statistiques descriptives sur le S&P 500 et les séries simulées avec les modèles

	μ [a]	σ [b]	$\tilde{\mu}_3$ [c]	$\tilde{\mu}_4$ [d]	Min	Max
S&P 500	0.018	1.256	−0.393	13.931	−12.765	10.957
HMM-Homogeneous($K = 2$)	0.013	1.199	−0.193	5.952	−9.282	9.997
HMM-Homogeneous($K = 3$)	0.013	1.198	−0.244	8.568	−14.334	12.154
HMM-Homogeneous($K = 4$)	0.013	1.205	−0.290	9.069	−12.510	14.234
HHMM($K = 4$)	0.013	1.201	−0.292	10.780	−14.532	14.169
HHMM-simplifié($K = 4$)	0.011	1.208	−0.313	9.601	−13.795	14.026
LOGIT($K = 4$) ∼ S&P 500[−1]	0.012	1.195	−0.157	8.549	−13.431	10.625
LOGIT($K = 4$) ∼ INDPROlog	0.021	1.036	−0.138	6.146	−13.443	11.471
DDMS($K = 4, \tau = 3$)	0.011	1.207	−0.193	8.438	−11.654	11.666
DDMS($K = 4, \tau = 5$)	0.011	1.205	−0.159	8.692	−11.471	12.407
DDMS($K = 4, \tau = 14$)	0.010	1.213	−0.263	8.480	−11.854	11.128

a. Moyenne inconditionnelle
b. Écart-type inconditionnel
c. Coefficient d'asymétrie : $\tilde{\mu}_3 = \mathbb{E}\left[\left(\frac{r-\mu}{\sigma}\right)^3\right]$
d. Coefficient d'aplatissement : $\tilde{\mu}_4 = \mathbb{E}\left[\left(\frac{r-\mu}{\sigma}\right)^4\right]$

6.2.5. Effet de levier

Les différences entre la corrélation des log-rendements centrés en valeur absolue avec un choc négatif et avec un choc positif (i.e. $\mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{-\epsilon_t, 0\}] - \mathbb{C}\text{orr}[|\epsilon_{t+h}|, \max\{\epsilon_t, 0\}]$) sur les séries simulées sont présentées à la figure 6.7 (p.122). La ligne bleue y représente cette différence observée sur la série du S&P 500.

Bien que cette différence ne soit pas significative pour la presque totalité des modèles, le modèle HMM-LOGIT($K = 4$) utilisant les valeurs précédentes de la série comme variable

explicative semble tout de même afficher une différence positive pour un délai relativement court. Ceci est potentiellement dû au fait que les plus grands coefficients de la matrice $\boldsymbol{\beta}_1$ sont négatifs, ce qui implique que les log-rendements négatifs ont un grand effet sur les probabilités de transition dans certains régimes. Ceci sera discuté plus en détails à la section 6.3.3 (p.125).

Il est de plus important de mentionner que bien que cette différence soit très faible (habituellement autour de 0.01), elle est toujours positive pour chacun de nos modèles.

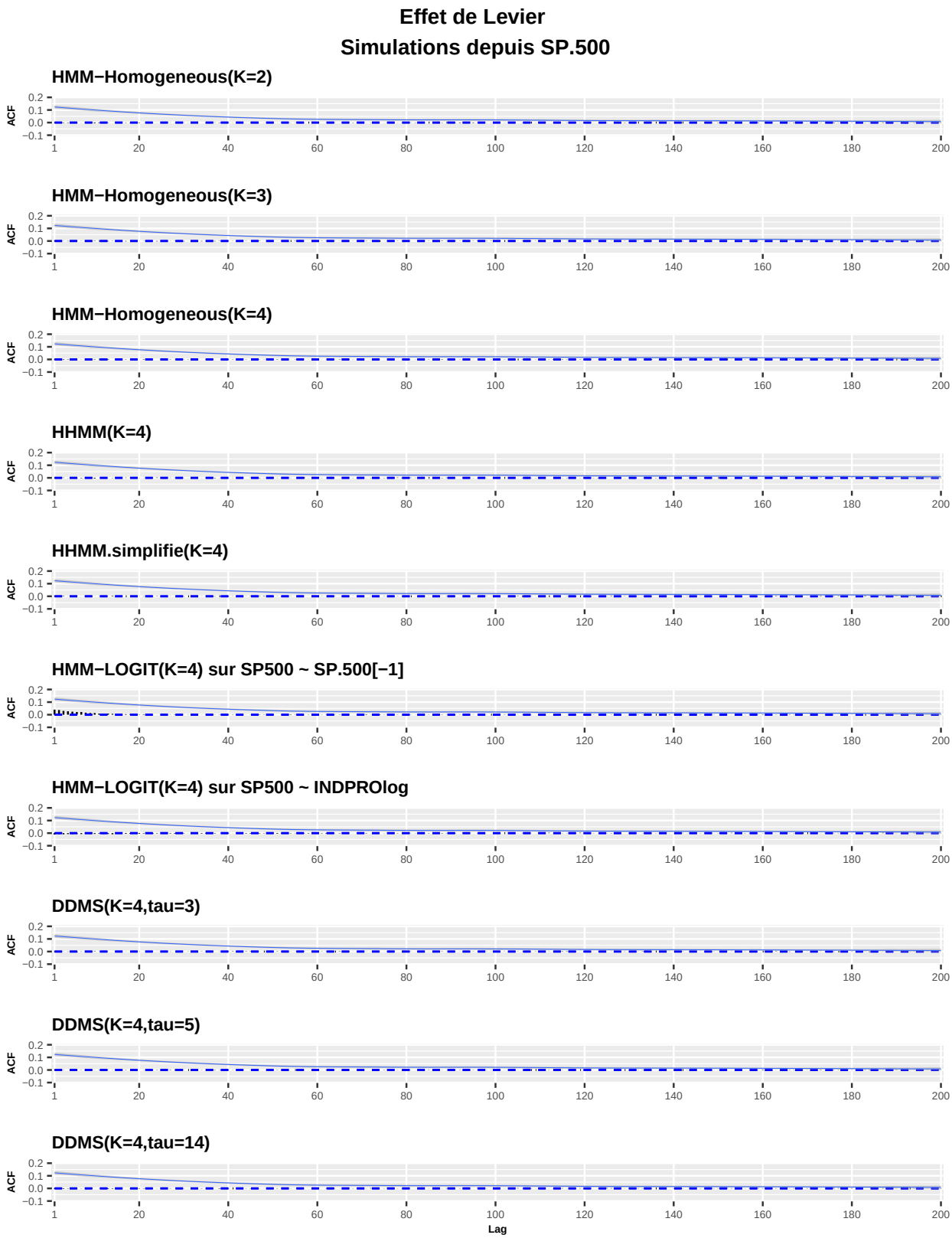


Figure 6.7 – Illustration de l'effet de levier sur le S&P 500 et les séries simulées avec les modèles à l'étude

6.3. Interprétation des paramètres estimés sur le S&P 500

6.3.1. Paramètres estimés pour le HHMM($K = 4$)

Les paramètres obtenus lors de l'estimation du HHMM($K = 4$) sont regroupés à la figure 6.8 (p.123) et la matrice de transition résultante est

$$\mathbf{\Gamma} = \begin{bmatrix} 0.965 & 0.035 & 0 & 0 \\ 0.006 & 0.977 & 0.018 & 0 \\ 0 & 0.010 & 0.953 & 0.037 \\ 0 & 0.001 & 0.031 & 0.968 \end{bmatrix}. \tag{6.3.1}$$

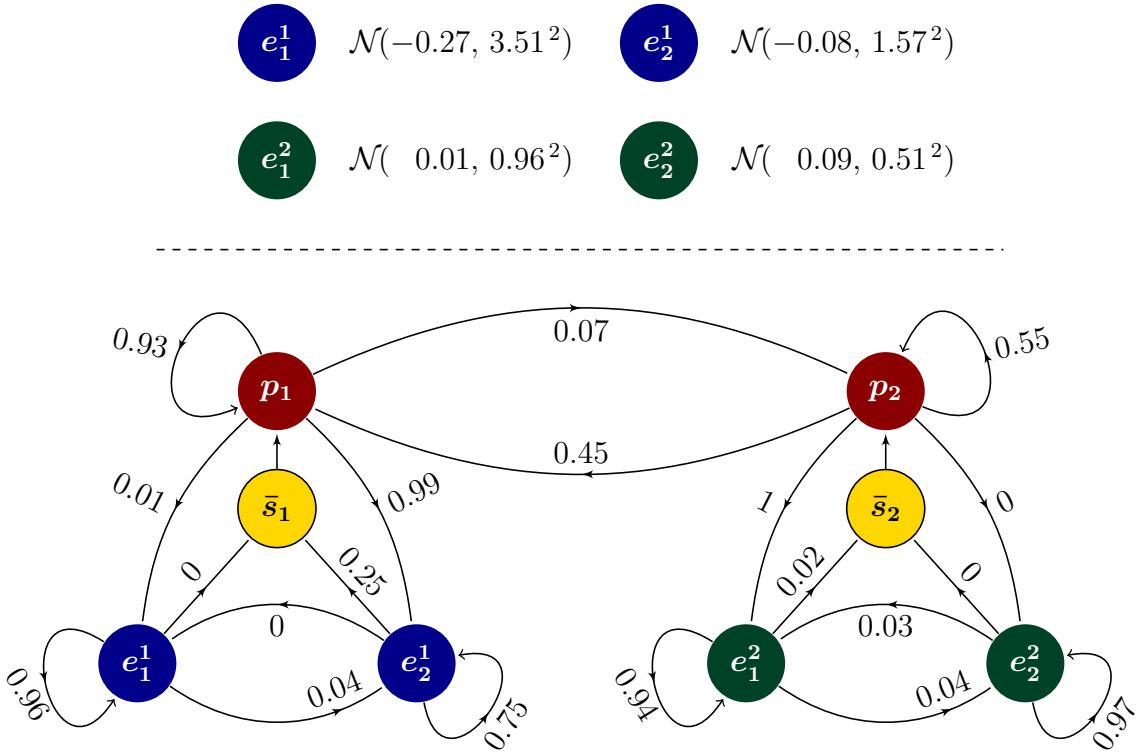


Figure 6.8 – Paramètres pour le HHMM($K = 4$) estimés sur le S&P 500

Au premier abord, il semble que les régimes soient bel et bien séparés en deux groupes distincts associés aux régimes de la couche supérieure de la chaîne. Les régimes *enfants* associés au régime *parent* p_1 semblent correspondre à des régimes plus turbulents avec une tendance à la baisse. En contrepartie, les régimes *enfants* associés au régime *parent* p_2 semblent associés à des états plus calmes de l'économie, avec une faible tendance à la hausse. Une analyse des

probabilités de transition indique aussi que le passage de l'état le plus calme à l'état le plus turbulent se fait toujours par le biais des états intermédiaires, avec une faible probabilité de quitter les états calmes, et une faible probabilité de quitter les états turbulents une fois dans ceux-ci. Ceci est cohérent avec les phases du cycle économique.

6.3.2. Paramètres estimés pour le HHMM-simplifié($K = 4$)

Les paramètres obtenus lors de l'optimisation du HHMM-simplifié($K = 4$) sont regroupés à la figure 6.9 (p.124) et la matrice de transition résultante est

$$\mathbf{\Gamma} = \begin{bmatrix} 0.951 & 0.043 & 0.006 & 0 \\ 0.005 & 0.988 & 0 & 0.006 \\ 0.003 & 0 & 0.926 & 0.070 \\ 0 & 0.004 & 0.054 & 0.942 \end{bmatrix}. \tag{6.3.2}$$

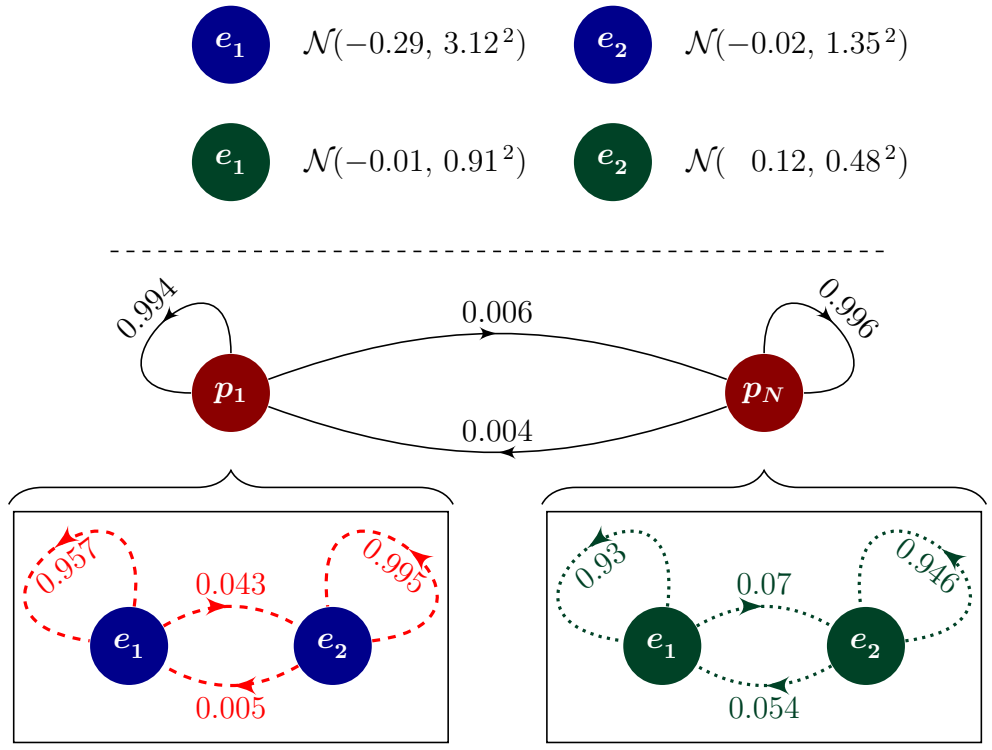


Figure 6.9 – Paramètres pour le HHMM-simplifié($K = 4$) estimés sur le S&P 500

Les paramètres estimés pour ce modèle semblent à nouveau se séparer en deux groupes distincts associés à des périodes plus turbulentes pour le régime *parent* p_1 et des périodes plus calmes associées au régime *parent* p_2. Sous cette spécification cependant, la probabilité

de passer d'un état intermédiaire à un autre est presque nulle. Nous observons sous cette spécification une grande persistance des régimes parents comparativement à la spécification générale présentée ci-haut. Ceci est potentiellement la raison pour laquelle le regroupement des volatilités semble plus clair pour cette spécification.

6.3.3. Paramètres estimés pour le HMM-LOGIT($K = 4$) $\sim$ S&P 500

Les paramètres des lois normales obtenus lors de l'estimation du HMM-LOGIT($K = 4$) utilisant les valeurs décalées de la variable dépendante sont

$$\begin{aligned}
&\text{Régime 1} : \mathcal{N}(-0.53,\ 4.77^{\,2}),\\
&\text{Régime 2} : \mathcal{N}(\ 0.004,\ 2.08^{\,2}),\\
&\text{Régime 3} : \mathcal{N}(-0.07,\ 1.09^{\,2}),\\
&\text{Régime 4} : \mathcal{N}(\ \ 0.09,\ 0.56^{\,2})
\end{aligned}$$

et les paramètres de la matrice de transition sont construits pas à pas avec la relation

$$\boldsymbol{f}_t = \underbrace{\begin{bmatrix} 18.071 & -120.988 & -30.083 & - \\ - & 10.895 & 7.586 & 7.766 \\ - & 46.328 & 49.954 & -59.311 \\ - & -13.097 & 17.777 & 22.132 \end{bmatrix}}_{\boldsymbol{\beta}_0} + \underbrace{\begin{bmatrix} 1.007 & -23.528 & -4.308 & - \\ - & 1.872 & 2.323 & 2.121 \\ - & 7.190 & 7.946 & -14.468 \\ - & 4.161 & -2.114 & -0.990 \end{bmatrix}}_{\boldsymbol{\beta}_1} \times X_t. \tag{6.3.3}$$

Tel que mentionné à la section 6.2.5 (p.120), ce modèle est le seul permettant d'observer l'effet de levier sur les données simulées. Ceci est possiblement lié au fait que les coefficents négatifs de la matrice $\boldsymbol{\beta}_1$ sont de plus grande ampleur que les coefficients positifs. Ceci implique qu'un log-rendement négatif aura un effet relatif plus grand qu'un log-rendement positif sur la probabilité de quitter l'état actuel. Il est aussi intéressant de noter la faible taille des coefficients de la matrice $\boldsymbol{\beta}_1$ par rapport à ceux de la matrice $\boldsymbol{\beta}_0$, ce qui signifie que les mouvements de l'indice S&P 500 ont un effet sur les probabilités de transition, mais que celui-ci est relativement limité.

6.3.4. Paramètres estimés pour le HMM-LOGIT($K = 4$) $\sim$ INDPRO

Les paramètres des lois normales obtenus lors de l'estimation du HMM-LOGIT($K = 4$) utilisant les log-rendements mensuels du niveau de INDPRO sont

$$\text{Régime 1} : \mathcal{N}(-0.59,\, 4.49^2),$$
$$\text{Régime 2} : \mathcal{N}(-0.05,\, 1.14^2),$$
$$\text{Régime 3} : \mathcal{N}(\ \ 0.09,\, 0.54^2),$$
$$\text{Régime 4} : \mathcal{N}(-0.04,\, 2.05^2)$$

et les paramètres de la matrice de transition sont construits pas à pas avec la relation

$$\boldsymbol{f}_t = \underbrace{\begin{bmatrix} 5.005 & -118.439 & -30.083 & - \\ - & 24.191 & 20.693 & 19.106 \\ - & 44.211 & 47.625 & -54.778 \\ - & -13.097 & 18.033 & 21.876 \end{bmatrix}}_{\boldsymbol{\beta}_0} + \underbrace{\begin{bmatrix} 0.415 & -27.058 & -4.308 & - \\ - & 5.559 & 5.442 & 5.612 \\ - & 15.263 & 16.201 & -30.994 \\ - & 4.161 & -1.774 & -1.330 \end{bmatrix}}_{\boldsymbol{\beta}_1} \times \vec{Y}_t. \tag{6.3.4}$$

Les paramètres obtenus pour les distributions normales de ce modèles sont très semblables à ceux obtenus pour le modèle précédent. Il semble aussi que les probabilités de transition de ce modèle dépendent du mouvement de l'indice INDPRO, mais que cette dépendance soit à nouveau limitée par l'ampleur relative des paramètres de la matrice $\boldsymbol{\beta}_0$ par rapport à la matrice $\boldsymbol{\beta}_1$. Sous cette spécification, nous observons cependant une plus grande dépendance au mouvement de l'indice INDPRO dans les régimes moins turbulents. Ceci est potentiellement expliqué par le fait que cet indice est habituellement moins turbulent que le S&P 500, qu'il ne change qu'une fois par mois, et qu'il est un bon indicateur de la santé de l'économie américaine.

6.3.5. Paramètres estimés pour le HMM-GAS($K = 2$)

Les paramètres des lois normales obtenus lors de l'estimation du HMM-GAS($K = 2$) sont

$$\text{Régime 1} : \mathcal{N}(-0.12,\, 3.85^2)$$
$$\text{Régime 2} : \mathcal{N}(\ \ 0.07,\, 0.39^2)$$

et les paramètres de la matrice de transition sont construits pas à pas avec la relation

$$\boldsymbol{f}_t = \underbrace{\begin{bmatrix} 0.410 \\ 0.728 \end{bmatrix}}_{\boldsymbol{\omega}} + \underbrace{\begin{bmatrix} 0.808 & - \\ - & 0.181 \end{bmatrix}}_{\mathbf{A}} \times \mathbf{s}_t + \underbrace{\begin{bmatrix} 0.98 & - \\ - & 0.991 \end{bmatrix}}_{\mathbf{B}} \times \boldsymbol{f}_{t-1}. \tag{6.3.5}$$

L'un des régimes est à nouveau associé à une distribution ayant une forte volatilité et une tendance à la décroissance (*bearish regime*), et le second associé à une faible volatilité et une légère tendance à la hausse (*bullish regime*).

À l'instar de Bazzi, Blasques, Koopman et Lucas [2017], la matrice $\mathbf{B}$ obtenue est presque égale à la matrice identité. Ceci à pour effet d'engendrer une grande persistance dans les régimes, laquelle est observable à la figure A.8 (p.147). De plus, le régime moins turbulent est associé à un paramètre constant plus grand et une entrée de la matrice $\mathbf{A}$ plus faible, impliquant une moindre dépendance aux changements de contexte économique, alors que le régime turbulent est plus sensible à un changement de contexte mais moins probable a priori.

6.3.6. Paramètres estimés pour le HMM-DDMS$(K = 4, \tau = 14)$

Nous restreindrons notre analyse des paramètres obtenus pour les modèles HMM-DDMS$(K = 4, \tau)$ au modèle admettant une mémoire maximale de $\tau = 14$ puisque c'est celui qui, en moyenne, a le mieux performé à la fois dans l'échantillon et à l'extérieur de celui-ci, pour l'ajustement aux données, l'erreur de prévision et la reproduction des faits stylisés.

Les paramètres des lois normales obtenus lors de l'estimation du HMM-DDMS$(K = 4, \tau = 14)$ sont

$$\begin{aligned}
\text{Régime 1} &: \mathcal{N}(-0.15,\ 2.65^2) \\
\text{Régime 2} &: \mathcal{N}(-0.07,\ 1.26^2) \\
\text{Régime 3} &: \mathcal{N}(\ \ 0.09,\ 0.53^2) \\
\text{Régime 4} &: \mathcal{N}(\ \ 1.41,\ 0.13^2)
\end{aligned}$$

et les paramètres de la matrice de transition sont construits pas à pas avec la relation

$$\boldsymbol{f}_t = \underbrace{\begin{bmatrix} 0.685 & -5.778 & -2.086 & - \\ - & 4.420 & 5.042 & -6.723 \\ - & 8.412 & 8.062 & -1.731 \\ - & -4.540 & -5.310 & 12.451 \end{bmatrix}}_{\boldsymbol{\beta}_0} + \underbrace{\begin{bmatrix} 0.547 & -5.801 & 0.466 & - \\ - & -0.034 & -0.418 & -9.685 \\ - & 8.751 & 9.064 & -5.711 \\ - & -1.113 & 9.590 & -6.840 \end{bmatrix}}_{\boldsymbol{\beta}_1} \times D_t. \quad (6.3.6)$$

Tel que nous le mentionnions à la section 6.2.3 (p.117) concernant le fait stylisé (3), le régime 4 associé à une faible volatilité et une forte tendance à la hausse ne semble jamais visité dans la série. Bien que ceci soit potentiellement dû à une mauvaise initialisation des paramètres, plusieurs essais sont malheureusement restés infructueux. À la section 6.1 (p.107), nous mentionnions que les modèles de bases ne semblaient pas suggérer la présence de quatre régimes sur cet intervalle de données. Il est donc possible que ceci ne soit en fait qu'une conséquence de la longue mémoire de ces modèles. Le coefficient $\beta_{44}^1 = -6.840$ étant fortement négatif relatif aux autres, il n'est pas surprenant que le modèle n'y reste que très peu de temps, si même il y passe.

Il est tout de même intéressant d'analyser les paramètres β_{ii}^1 associés aux autres régimes afin d'étudier l'effet de la durée sur leur persistance. Il semble d'abord que le nombre de pas dans le régime le plus turbulent soit lié à une augmentation de sa persistance, mais que cet effet soit relativement faible. Pour ce qui est de l'état intermédiaire, il semble que celui-ci devienne de moins en moins persistant au fur et à mesure que nous nous enfonçons dans celui-ci. En contrepartie, il semble que l'état calme devienne de plus en plus persistant lorsque le nombre de pas dans celui-ci augmente. Nous pouvons aisément observer ces caractéristiques à la figure A.11 (p.150) dans laquelle les zones jaunes sont relativement courtes et éparses, les zones rouges généralement plus larges et les zones vertes légèrement plus larges encore, mais surtout plus fréquentes.

Conclusion

Nous avons présenté dans ce mémoire les modèles de mélange, les *HMMs* standards ainsi que deux reformulations hiérarchiques et trois extensions non-homogènes, à savoir le modèle HMM-LOGIT(K), le HMM-GAS(K) et finalement le HMM-DDMS(K,τ). Notre objectif était d'étudier l'apport potentiel de structures hiérarchiques et de la relaxation de l'hypothèse d'homogénéité à la qualité de l'ajustement aux données et à la reproduction des faits stylisés grâce aux modèles de Markov à variables latentes. Nous étions de plus intéressés par l'interprétabilité des paramètres des divers modèles étudiés.

À la lumière des analyses présentées au chapitre 6, il semble que la reformulation hiérarchique et la relaxation de l'hypothèse d'homogénéité permettent aux modèles de reproduire plus fidèlement les faits stylisés. Bien que ce soit les modèles hiérarchiques qui prennent la première place pour la majorité des méthodes de sélection de modèles, l'apport des modèles non-homogènes n'est cependant pas négligeable. En effet, la qualité des prévisions obtenues sur le S&P 500 grâce au modèle HMM-LOGIT($K = 4$) est nettement supérieure à tous les autres modèles et l'inclusion de valeurs décalées de la variable à l'étude dans ce dernier semble permettre une meilleure reproduction de l'effet de levier. Pour leur part, les modèles HMM-GAS($K = 2$) se démarquent par leur capacité à capter les dynamiques du S&P 500 et la rapidité avec laquelle ils convergent. L'inclusion de la durée comme variable explicative semble de plus permettre de reproduire plus fidèlement la lente décroissance de l'autocorrélation des processus $\{|\epsilon_t|\}$ et $\{{\epsilon_t}^2\}$, fait stylisé particulièrement difficile à reproduire selon Rydén, Teräsvirta et Åsbrink [1998]. De plus, malgré la complexité croissante des modèles présentés, leurs paramètres peuvent malgré tout être interprétés et liés à des caractéristiques observables sur la série à l'étude.

Ainsi, il serait intéressant d'étendre l'étude actuelle en combinant dans un premier temps une approche hiérarchique avec une dynamique non-homogène. En particulier, nous croyons que l'inclusion de variables explicatives à la manière des HMM-LOGIT(K) et HMM-GAS(K) pourrait permettre de combiner la faculté des *HHMMs* à reproduire les faits stylisés avec la faculté desdits modèles à faire des prévisions sur le S&P 500. L'utilisation de valeurs décalées de la variable dépendante dans une structure analogue au HMM-LOGIT(K) permetterait

potentiellement aussi au modèle résultant de reproduire l'effet de levier, complétant pour ainsi dire la liste des faits stylisés reproductibles par les modèles hiérarchiques.

Remarque 6.3.1. Puisque la reformulation hiérarchique inclut plusieurs couches de chaînes de Markov, l'utilisation de celle-ci ouvre de plus la porte à une panoplie de nouvelles formulations. Il serait à cet égard intéressant de tester l'inclusion de paramètres variables dans une seule couche, ou d'inclure des séries différentes comme variables explicatives dans les différentes couches. Ceci permettrait par exemple d'inclure une série plus stable représentant l'état général de l'économie dans la couche supérieure et une série propre au même secteur de l'économie que la série étudiée dans une couche subalterne. L'apport d'une telle distinction entre les couches est cependant spéculatif, il conviendrait d'abord de tester l'inclusion de plus d'une série de manière plus généralisée et d'analyser les paramètres obtenus.

Dans un second temps, considérant la performance remarquable des modèles hiérarchiques, il serait intéressant de comparer ces derniers avec les modèles factoriels mentionnés en introduction du chapitre 4. Ne représentant en fait qu'une différence au niveau de la transformation appliquée à la série d'états générée par les couches successives, ceux-ci pourraient s'inscrire dans une étude plus large sur les diverses transformations proposées dans la littérature.

Finalement, bien que nous n'ayons pas obtenu de bons résultats avec les *DDMSs* en terme de log-vraisemblance et de prévision, ils semblaient reproduire les faits stylisés relativement bien. Puisqu'ils partageaient la même relation entre les paramètres et les probabilités de transition, mais se démarquaient pour des mesures de performance différentes, il serait intéressant d'intégrer la durée provenant du modèle HMM-DDMS(K,τ) aux côtés de séries financières dans un modèle de type HMM-LOGIT(K). Les *DDMSs* se démarquant de plus des autres modèles non-homogènes par le fait qu'il est possible de les reformuler comme des *HMMs* homogènes, ils se présentent à notre avis comme des candidats particulièrement intéressants pour ce type de combinaison.

Bibliographie

Ang, Andrew et Geert Bekaert (2002a). "International Asset Allocation With Regime Shifts". In : *Review of Financial Studies* 15.4, p. 1137-1187. URL : https://EconPapers.repec.org/RePEc:oup:rfinst:v:15:y:2002:i:4:p:1137-1187.

— (2002b). "Regime Switches in Interest Rates". In : *Journal of Business & Economic Statistics* 20.2, p. 163-182. DOI : 10.1198/073500102317351930. eprint : https://doi.org/10.1198/073500102317351930. URL : https://doi.org/10.1198/073500102317351930.

— (2004). "How Regimes Affect Asset Allocation". In : *Financial Analysts Journal* 60.2, p. 86-99. ISSN : 0015198X. URL : http://www.jstor.org/stable/4480559.

Ang, Andrew et Joseph Chen (2002). "Asymmetric correlations of equity portfolios". In : *Journal of Financial Economics* 63.3, p. 443 -494. ISSN : 0304-405X. DOI : https://doi.org/10.1016/S0304-405X(02)00068-5. URL : http://www.sciencedirect.com/science/article/pii/S0304405X02000685.

Ang, Andrew et Allan Timmermann (2012). "Regime Changes and Financial Markets". In : *Annual Review of Financial Economics* 4.1, p. 313-337. URL : https://EconPapers.repec.org/RePEc:anr:refeco:v:4:y:2012:p:313-337.

Augustyniak, M. (jan. 2018). *STT6705V : Méthodes d'inférence pour les modèles à chaîne de Markov cachée [Notes de cours]*.

Augustyniak, Maciej (2014). "Maximum likelihood estimation of the Markov-switching GARCH model". In : *Computational Statistics & Data Analysis* 76. CFEnetwork : The Annals of Computational and Financial Econometrics, p. 61-75. ISSN : 0167-9473. DOI : https://doi.org/10.1016/j.csda.2013.01.026. URL : https://www.sciencedirect.com/science/article/pii/S0167947313000418.

Augustyniak, Maciej, Luc Bauwens et Arnaud Dufays (fév. 2019). "A New Approach to Volatility Modeling : The Factorial Hidden Markov Volatility Model". In : *Journal of Business & Economic Statistics*. DOI : 10.1080/07350015.2017.1415910.

Baum, Leonard E. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes". In : *Inequalities*

III : Proceedings of the Third Symposium on Inequalities. Sous la dir. d'Oved Shisha. University of California, Los Angeles : Academic Press, p. 1-8.

Baum, Leonard E. et J. A. Eagon (mai 1967). "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology". In : *Bulletin of the American Mathematical Society* 73.3, p. 360-364. doi : 10.1090/s0002-9904-1967-11751-8. url : https://doi.org/10.1090%2Fs0002-9904-1967-11751-8.

Baum, Leonard E. et Ted Petrie (déc. 1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". In : *Ann. Math. Statist.* 37.6, p. 1554-1563. doi : 10.1214/aoms/1177699147. url : https://doi.org/10.1214/aoms/1177699147.

Baum, Leonard E., Ted Petrie, George Soules et Norman Weiss (fév. 1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains". In : *Ann. Math. Statist.* 41.1, p. 164-171. doi : 10.1214/aoms/1177697196. url : https://doi.org/10.1214/aoms/1177697196.

Baum, Leonard E. et George R. Sell (1968). "Growth transformations for functions on manifolds." In : *Pacific J. Math.* 27.2, p. 211-227. url : https://projecteuclid.org:443/euclid.pjm/1102983899.

Bazzi, Marco (2017). *HMMGAS : Hidden Markov Models with Generalized Autoregressive Score*. R package version 2.0.

Bazzi, Marco, Francisco Blasques, Siem Jan Koopman et Andre Lucas (mai 2017). "Time-Varying Transition Probabilities for Markov Regime Switching Models". In : *Journal of Time Series Analysis* 38.3, p. 458-478. url : https://ideas.repec.org/a/bla/jtsera/v38y2017i3p458-478.html.

Bengio, Y. (juill. 1999). "Markovian Models for Sequential Data". In : *Neural Computing Surveys* 2.

Bengio, Yoshua et Paolo Frasconi (1994). "An Input Output HMM Architecture". In : *Proceedings of the 7th International Conference on Neural Information Processing Systems*. NIPS'94. Cambridge, MA, USA : MIT Press, 427–434.

Bergstra, James et Yoshua Bengio (fév. 2012). "Random Search for Hyper-Parameter Optimization". In : *J. Mach. Learn. Res.* 13.null, 281–305. issn : 1532-4435.

Boothe, Paul et Debra Glassman (1987). "The statistical distribution of exchange rates : Empirical evidence and economic implications". In : *Journal of International Economics* 22.3-4, p. 297-319.

Bulla, Jan (2011). "Hidden Markov models with t components. Increased persistence and other aspects". In : *Quantitative Finance* 11.3, p. 459-475. doi : 10.1080/14697681003685563. eprint : https://doi.org/10.1080/14697681003685563. url : https://doi.org/10.1080/14697681003685563.

Bulla, Jan et Ingo Bulla (2006). "Stylized facts of financial time series and hidden semi-Markov models". In : *Computational Statistics & Data Analysis* 51.4. Nonlinear Modelling

and Financial Econometrics, p. 2192 -2209. ISSN : 0167-9473. DOI : https://doi.org/10.1016/j.csda.2006.07.021. URL : http://www.sciencedirect.com/science/article/pii/S0167947306002374.

BULLA, Jan, Sascha MERGNER, Ingo BULLA, André SESBOÜÉ et Christophe CHESNEAU (juill. 2011). "Markov-switching asset allocation : Do profitable strategies exist ?" In : *Journal of Asset Management* 12, p. 310-321. DOI : 10.1057/jam.2010.27.

CALVET, Laurent et Adlai J. FISHER (fév. 2004). "How to Forecast Long-Run Volatility : Regime Switching and the Estimation of Multifractal Processes". In : *Journal of Financial Econometrics* 2, p. 49-83. DOI : 10.1093/jjfinec/nbh003.

CHARLOT, Philippe et Vêlayoudom MARIMOUTOU (2011). *Hierarchical hidden Markov structure for dynamic correlations : the hierarchical RSDC model (version révisée)*. Working Papers hal-00605965. HAL. URL : https://ideas.repec.org/p/hal/wpaper/hal-00605965.html.

COCULESCU, Delia et Ashkan NIKEGHBALI (2007). *Filtrations*. arXiv : 0712.0622 [math.PR].

COX, D. R., Gudmundur GUDMUNDSSON, Georg LINDGREN, Lennart BONDESSON, Erik HARSAAE, Petter LAAKE, Katarina JUSELIUS et Steffen L. LAURITZEN (1981). "Statistical Analysis of Time Series : Some Recent Developments [with Discussion and Reply]". In : *Scandinavian Journal of Statistics* 8.2, p. 93-115. ISSN : 03036898, 14679469. URL : http://www.jstor.org/stable/4615819.

CREAL, Drew, Siem Jan KOOPMAN et Andre LUCAS (2013). "GENERALIZED AUTOREGRESSIVE SCORE MODELS WITH APPLICATIONS". In : *Journal of Applied Econometrics* 28.5, p. 777-795. URL : https://EconPapers.repec.org/RePEc:wly:japmet:v:28:y:2013:i:5:p:777-795.

CREAL, Drew, Siem Jan KOOPMAN et André LUCAS (2011). "A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations". In : *Journal of Business & Economic Statistics* 29.4, p. 552-563. DOI : 10.1198/jbes.2011.10070. eprint : https://doi.org/10.1198/jbes.2011.10070. URL : https://doi.org/10.1198/jbes.2011.10070.

CREAL, Drew D., S. J. KOOPMAN et A. LUCAS (2008). "A General Framework for Observation Driven Time-Varying Parameter Models". In : *Econometrics : Multiple Equation Models eJournal*.

DEMPSTER, A. P., N. M. LAIRD et D. B. RUBIN (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In : *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, p. 1-38. ISSN : 00359246. URL : http://www.jstor.org/stable/2984875.

Diebold, Francis, Joon-Haeng Lee et Gretchen C. Weinbach (1994). *Regime switching with time-varying transition probabilities*. Working Papers 93-12. Federal Reserve Bank of Philadelphia. URL : https://EconPapers.repec.org/RePEc:fip:fedpwp:93-12.

Diebold, F.X. (1988). *Empirical Modeling of Exchange Rate Dynamics*. Lecture Notes in Mathematics. Springer-Verlag. ISBN : 9783540189664. URL : https://books.google.ca/books?id=Jn6wAAAAIAAJ.

Doob, J.L. (1953). *Stochastic processes*. Wiley publications in statistics. Wiley. URL : https://books.google.ca/books?id=0GUGAQAAIAAJ.

Doornik, Jurgen (fév. 2013). "A Markov-switching model with component structure for US GNP". In : *Economics Letters* 118, 265–268. DOI : 10.1016/j.econlet.2012.10.035.

Durland, J Michael et Tom McCurdy (1994). "Duration-Dependent Transitions in a Markov Model of U.S. GNP Growth". In : *Journal of Business & Economic Statistics* 12.3, p. 279-88. URL : https://EconPapers.repec.org/RePEc:bes:jnlbes:v:12:y:1994:i:3:p:279-88.

Engel, Charles et James Hamilton (fév. 1990). "Long Swings in the Dollar : Are They in the Data and Do Markets Know It ?" In : *American Economic Review* 80, p. 689-713.

Ephraim, Yariv et Neri Merhav (juill. 2002). "Hidden Markov Processes". In : *Information Theory, IEEE Transactions on* 48, p. 1518 -1569. DOI : 10.1109/TIT.2002.1003838.

Fama, Eugene F. (1965). "The Behavior of Stock-Market Prices". In : *The Journal of Business* 38.1, p. 34-105. ISSN : 00219398, 15375374. URL : http://www.jstor.org/stable/2350752.

Federal Reserve System (US), Board of Governors of the (déc. 2020). *Industrial Production : Total Index [INDPRO]*. données recuillies depuis la FRED, Federal Reserve Bank of St. Louis ; https://fred.stlouisfed.org/series/INDPRO.

Filardo, Andrew J. (1994). "Business-Cycle Phases and Their Transitional Dynamics". In : *Journal of Business & Economic Statistics* 12.3, p. 299-308. ISSN : 07350015. URL : http://www.jstor.org/stable/1392086.

Fine, Shai, Yoram Singer et Naftali Tishby (1998). "The Hierarchical Hidden Markov Model : Analysis and Applications". In : *Machine Learning* 32.1, p. 41-62. DOI : 10.1023/A:1007469218079. URL : https://doi.org/10.1023/A:1007469218079.

Granger, Clive et Zhuanxin Ding (1995). "Some Properties of Absolute Return : An Alternative Measure of Risk". In : *Annals of Economics and Statistics* 40, p. 67-91. URL : https://EconPapers.repec.org/RePEc:adr:anecst:y:1995:i:40:p:67-91.

Hamilton, James (jan. 2008). "Regime-Switching Models". In : DOI : 10.1057/9780230280830_23.

Hamilton, James D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle". In : *Econometrica* 57.2, p. 357-384. ISSN : 00129682, 14680262. URL : http://www.jstor.org/stable/1912559.

— (jan. 1994). *Time Series Analysis*. 1re éd. Princeton University Press. ISBN : 0691042896. URL : http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0691042896.

Juang, B. H. et L. R. Rabiner (1991). "Hidden Markov Models for Speech Recognition". In : *Technometrics* 33.3, p. 251-272. ISSN : 00401706. URL : http://www.jstor.org/stable/1268779.

Kim, Chang-Jin et Charles Nelson (fév. 1998). "Business Cycle Turning Points, A New Coincident Index, And Tests Of Duration Dependence Based On A Dynamic Factor Model With Regime Switching". In : *The Review of Economics and Statistics* 80, p. 188-201. DOI : 10.1162/003465398557447.

Knuth, Donald E. (1984). *The TEX Book*. 15th. Addison-Wesley.

Kritzman, Mark, Sébastien Page et David Turkington (2012). "Regime Shifts : Implications for Dynamic Strategies (corrected)". In : *Financial Analysts Journal* 68.3, p. 22-39. DOI : 10.2469/faj.v68.n3.3. eprint : https://doi.org/10.2469/faj.v68.n3.3. URL : https://doi.org/10.2469/faj.v68.n3.3.

Lam, Pok san (1997). *A Markov switching model of GNP growth with duration dependence*. Discussion Paper / Institute for Empirical Macroeconomics 124. Federal Reserve Bank of Minneapolis. URL : https://ideas.repec.org/p/fip/fedmem/124.html.

Lamport, Leslie (1986). *LATEX : A Document Preparation System*. Reading, MA : Addison–Wesley Pub. Co.

Langouët, Hoël (juin 2011). "Constraints Derivative-Free Optimization : two industrial applications in reservoir engineering and in engine calibration". Theses. Université Nice Sophia Antipolis. URL : https://tel.archives-ouvertes.fr/tel-00671987.

Le, Nhu D., Brian G. Leroux, Martin L. Puterman et Paul S. Albert (1992). "Exact Likelihood Evaluation in a Markov Mixture Model for Time Series of Seizure Counts". In : *Biometrics* 48.1, p. 317-323. ISSN : 0006341X, 15410420. URL : http://www.jstor.org/stable/2532758.

Lee, Joon-Haneg (1991). "Nonstationary Markov Switching Models of Exchange Rates : The Pound-Dollar Exchange Rate". Thèse de doct. University of Pensilvania.

Liu, Xinyi, Dimitris Margaritis et Peiming Wang (2012). "Stock market volatility and equity returns : Evidence from a two-state Markov-switching model with regressors". In : *Journal of Empirical Finance* 19.4, p. 483 -496. ISSN : 0927-5398. DOI : https://doi.org/10.1016/j.jempfin.2012.04.011. URL : http://www.sciencedirect.com/science/article/pii/S0927539812000382.

Lystig, Theodore C et James P Hughes (2002). "Exact Computation of the Observed Information Matrix for Hidden Markov Models". In : *Journal of Computational and Graphical Statistics* 11.3, p. 678-689. DOI : 10.1198/106186002402. eprint : https://doi.org/10.1198/106186002402. URL : https://doi.org/10.1198/106186002402.

Maheu, John M. et Thomas H. McCurdy (2000a). "Identifying Bull and Bear Markets in Stock Returns". In : *Journal of Business & Economic Statistics* 18.1, p. 100-112. ISSN : 07350015. URL : http://www.jstor.org/stable/1392140.

— (2000b). "Volatility dynamics under duration-dependent mixing". In : *Journal of Empirical Finance* 7.3. Special issue on Risk Management, p. 345 -372. ISSN : 0927-5398. DOI : https://doi.org/10.1016/S0927-5398(00)00018-9. URL : http://www.sciencedirect.com/science/article/pii/S0927539800000189.

(MSFT), Microsoft Corporation (déc. 2020a). *GSPC*. données recuillies sur Yahoo Finance, https://finance.yahoo.com/quote/%5EGSPC?p=%5EGSPC.

— (déc. 2020b). *USD/CAD (CAD=X)*. données recuillies sur Yahoo Finance, https://finance.yahoo.com/quote/CAD%3DX?p=CAD%3DX.

Nystrup, Peter, Bo William Hansen, Henrik Olejasz Larsen, Henrik Madsen et Erik Lindström (2017). "Dynamic Allocation or Diversification : A Regime-Based Approach to Multiple Assets". In : *The Journal of Portfolio Management* 44.2, p. 62-73. ISSN : 0095-4918. DOI : 10.3905/jpm.2018.44.2.062. eprint : https://jpm.pm-research.com/content/44/2/62.full.pdf. URL : https://jpm.pm-research.com/content/44/2/62.

Nystrup, Peter, Bo William Hansen, Henrik Madsen et Erik Lindström (2015). "Regime-Based Versus Static Asset Allocation : Letting the Data Speak". In : *The Journal of Portfolio Management* 42.1, p. 103-109. ISSN : 0095-4918. DOI : 10.3905/jpm.2015.42.1.103. eprint : https://jpm.pm-research.com/content/42/1/103.full.pdf. URL : https://jpm.pm-research.com/content/42/1/103.

Nystrup, Peter, Henrik Madsen et Erik Lindström (2017). "Long Memory of Financial Time Series and Hidden Markov Models with Time-Varying Parameters". In : *Journal of Forecasting* 36.8, p. 989-1002. DOI : https://doi.org/10.1002/for.2447. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2447. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2447.

Osborne, M. F. M. (1959). "Brownian Motion in the Stock Market". In : *Operations Research* 7.2, p. 145-173. ISSN : 0030364X, 15265463. URL : http://www.jstor.org/stable/167153.

Perez-Quiros, Gabriel et Allan Timmermann (2000). "Firm Size and Cyclical Variations in Stock Returns". In : *Journal of Finance* 55.3, p. 1229-1262. URL : https://EconPapers.repec.org/RePEc:bla:jfinan:v:55:y:2000:i:3:p:1229-1262.

R Core Team (2013). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL : http://www.R-project.org/.

Rabiner, L. R. (fév. 1989). “A tutorial on hidden Markov models and selected applications in speech recognition”. In : *Proceedings of the IEEE* 77.2, p. 257-286. ISSN : 1558-2256. DOI : 10.1109/5.18626.

Robertson, Andrew, Sergey Kirshner et Padhraic Smyth (déc. 2003). “Hidden Markov models for modeling daily rainfall occurrence over Brazil”. In.

Rossi, Alessandro et Giampiero Gallo (juin 2002). *Volatility Estimation via Hidden Markov Models*. Econometrics Working Papers Archive. Universita’ degli Studi di Firenze, Dipartimento di Statistica Informatica Applicazioni "G. Parenti". URL : https://EconPapers.repec.org/RePEc:fir:econom:wp2002_14.

Rydén, Tobias, Timo Teräsvirta et Stefan Åsbrink (1998). “Stylized Facts of Daily Return Series and the Hidden Markov Model”. In : *Journal of Applied Econometrics* 13.3, p. 217-244. ISSN : 08837252, 10991255. URL : http://www.jstor.org/stable/223228.

Tan, Chia, you beng Koh, Kok Haur Ng et Kooi Ng (jan. 2020). “Dynamic Volatility Modelling of Bitcoin Using Time-Varying Transition Probability Markov-Switching GARCH Model”. In : *SSRN Electronic Journal*. DOI : 10.2139/ssrn.3573537.

Taylor, Stephen J (1986). *Modelling Financial Time Series*. 1st. Wiley et Sons, Incorporated, John. ISBN : 9780471909934.

TSX (2020). *TMX TSX | TSXV - Calendar*. [En ligne ; Page modifiée le 17 mars 2020 à 17 :14 :09 HAE]. URL : \url{https://www.tsx.com/trading/calendars-and-trading-hours/calendar}.

Wikipedia contributors (2019). *Trading day — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Trading_day&oldid=932281882. [Online ; accessed 17-March-2020].

Yoon, Byung-Jun (sept. 2009). “Hidden Markov Models and their Applications in Biological Sequence Analysis”. In : *Current genomics* 10, p. 402-15. DOI : 10.2174/138920209789177575.

Zucchini, Walter et Iain Macdonald (avr. 2009). *Hidden Markov Models for Time Series : An Introduction Using R*. ISBN : 9781420010893. DOI : 10.1201/9781420010893.

Annexe A

Graphiques et tableaux de résultats supplémentaires

A.1. Résultats pour la série S&P500

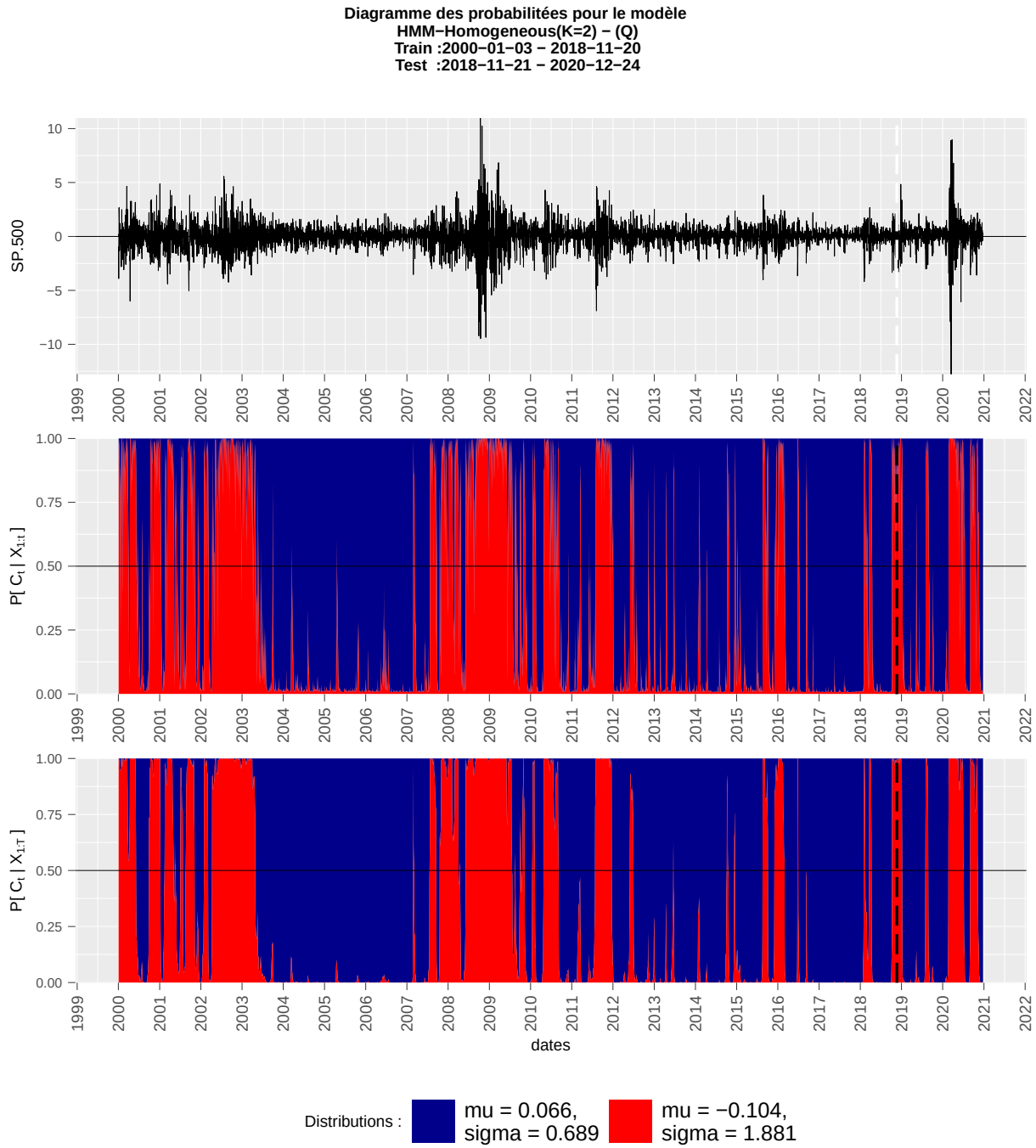


Figure A.1 – Probabilités filtrées et lissées pour le modèle HMM($K = 2$) estimées sur le S&P500

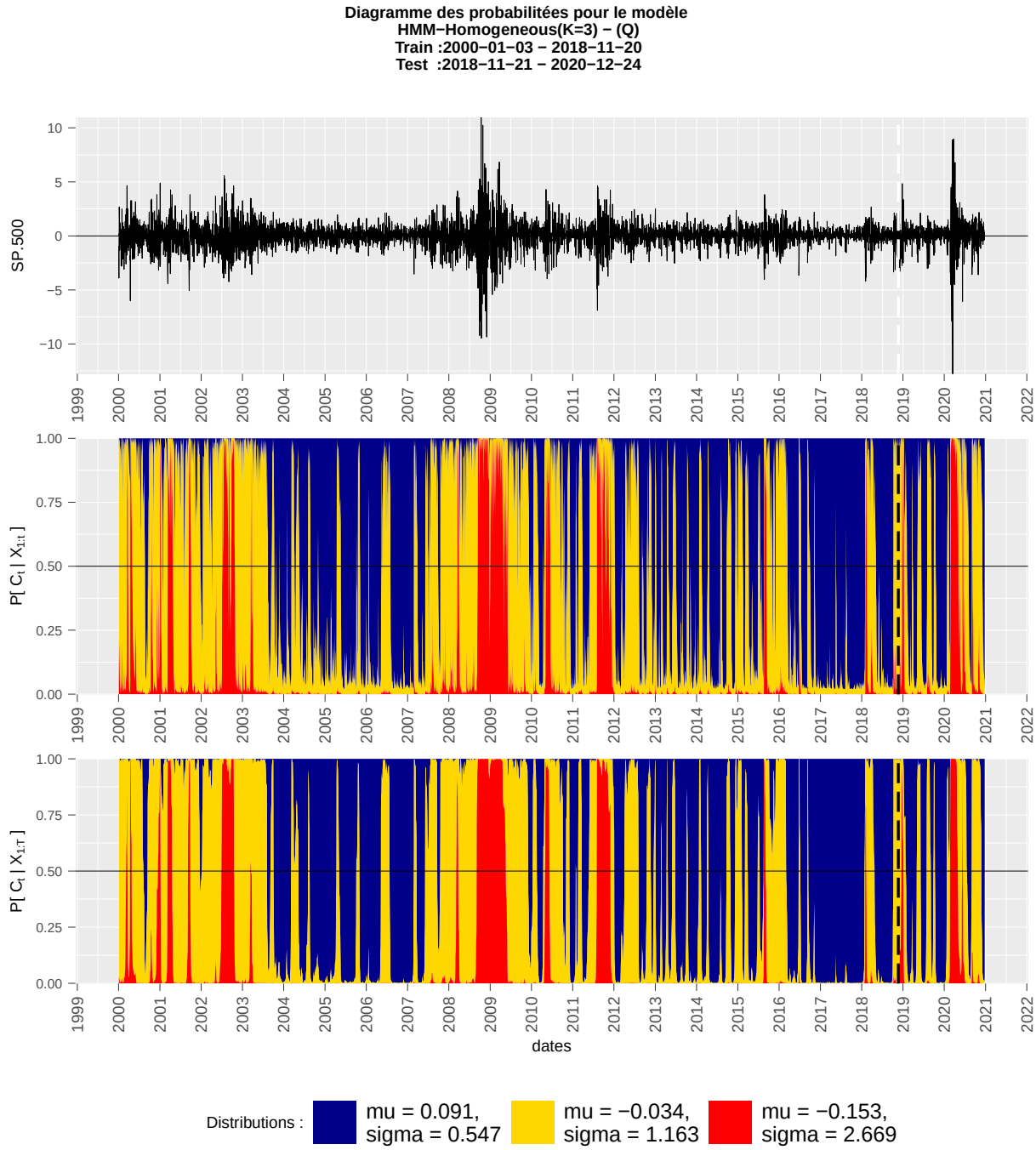


Figure A.2 – Probabilités filtrées et lissées pour le modèle HMM($K = 3$) estimées sur le S&P500

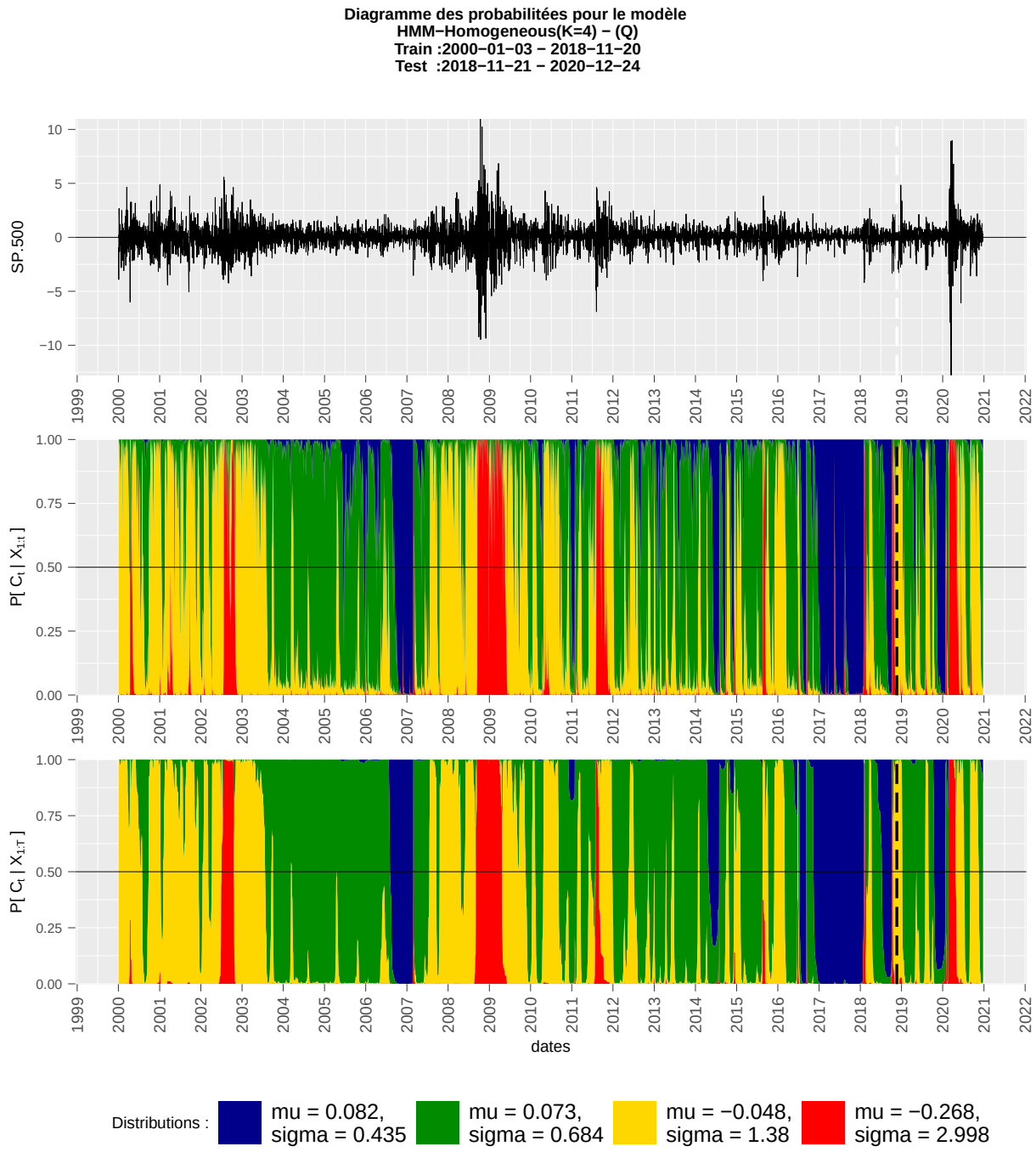


Figure A.3 – Probabilités filtrées et lissées pour le modèle HMM($K = 4$) estimées sur le S&P500

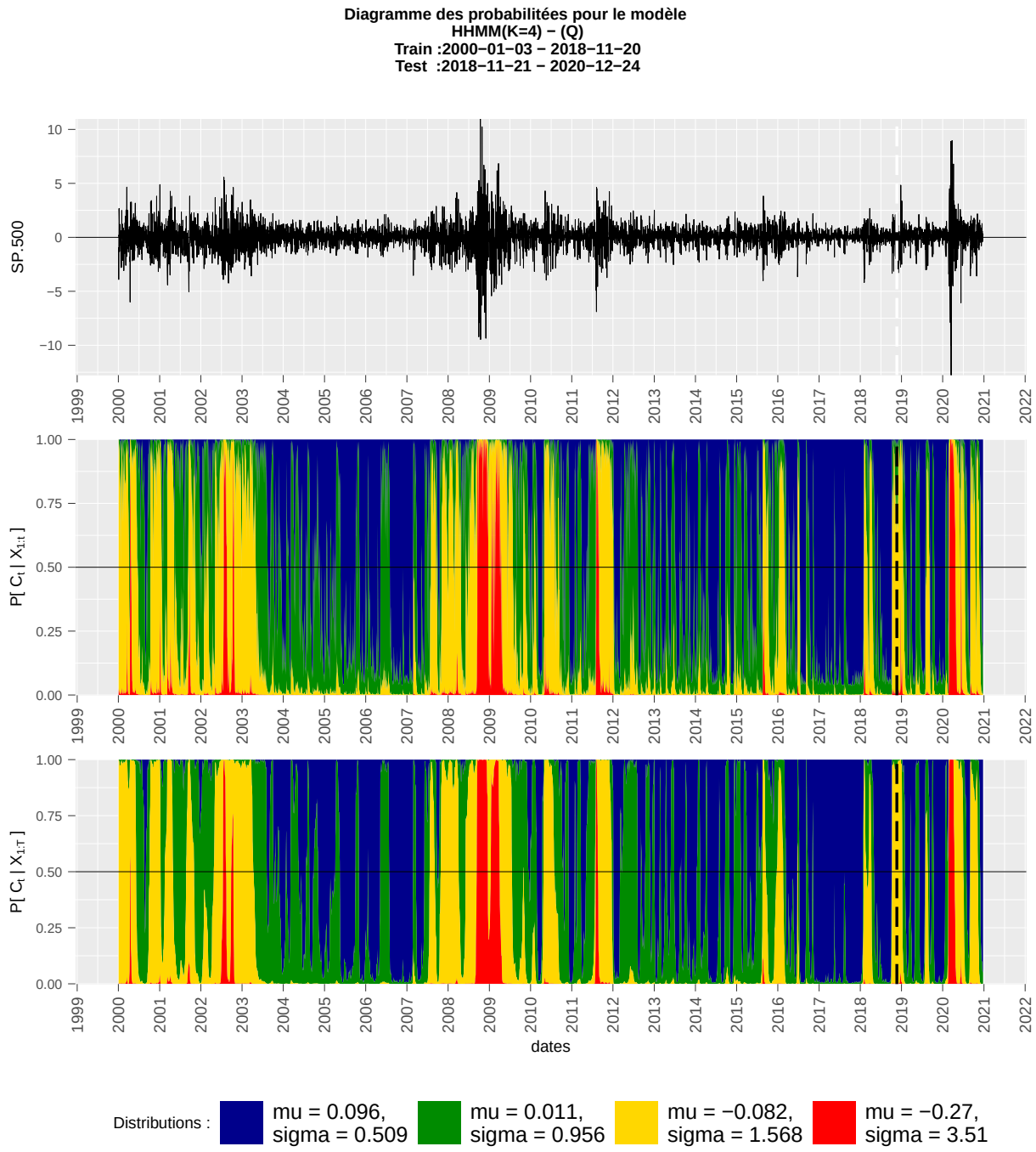


Figure A.4 – Probabilités filtrées et lissées pour le modèle HHMM($K = 4$) estimées sur le S&P 500

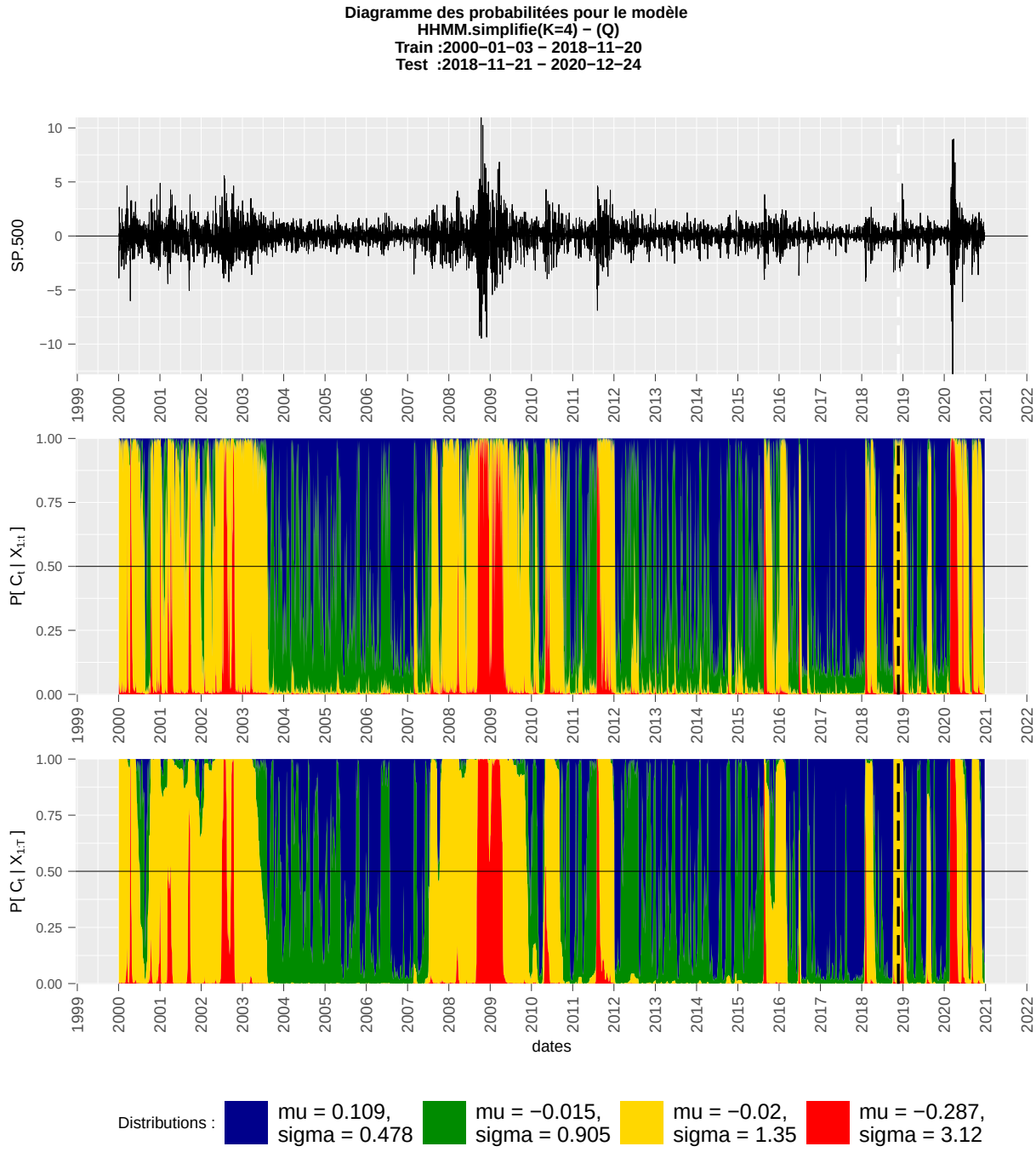


Figure A.5 – Probabilités filtrées et lissées pour le modèle HHMM-simplifié($K = 4$) estimées sur le S&P 500

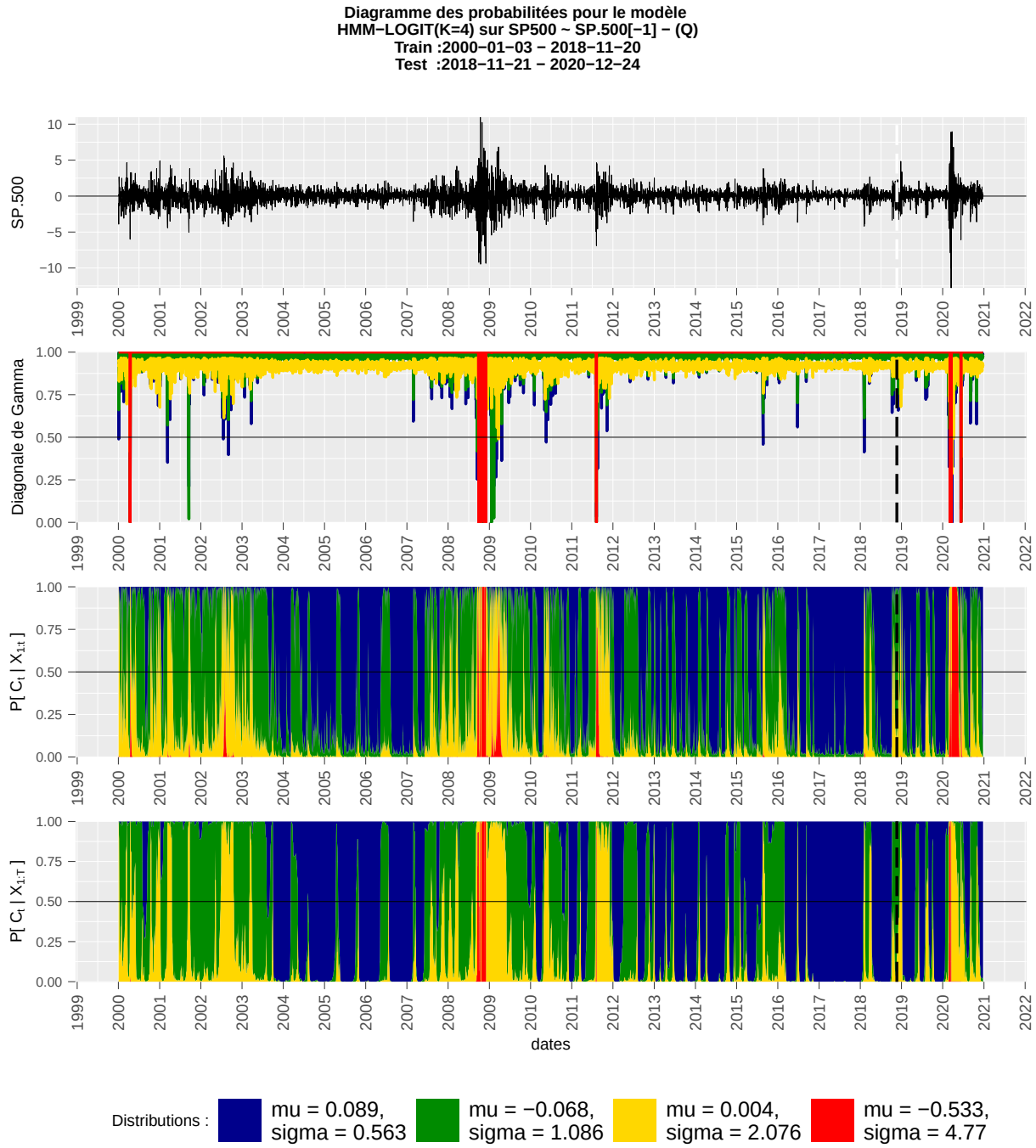


Figure A.6 – Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le S&P 500 avec ses valeurs décalées comme variable explicative

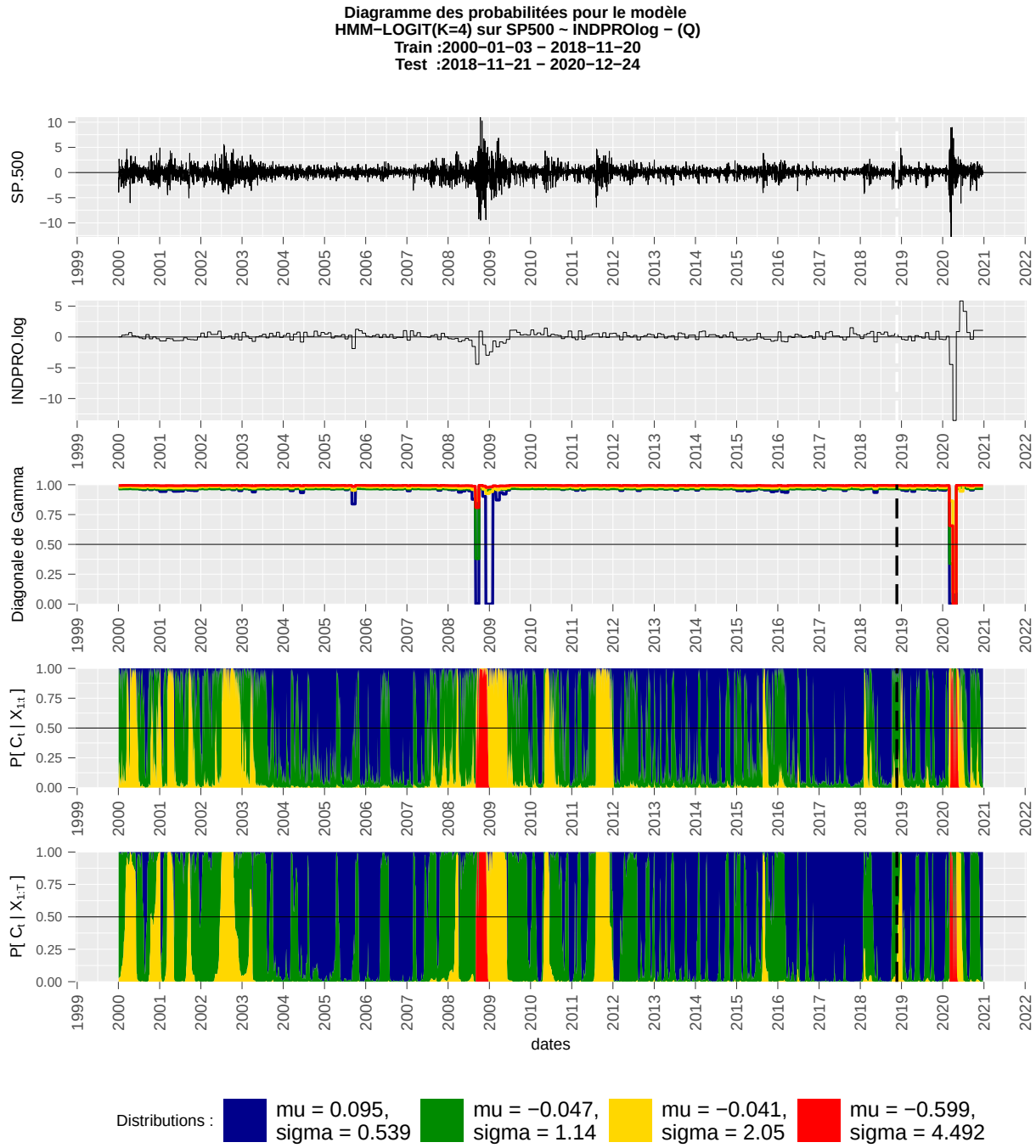


Figure A.7 – Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le S&P 500 avec la série INDPRO comme variable explicative

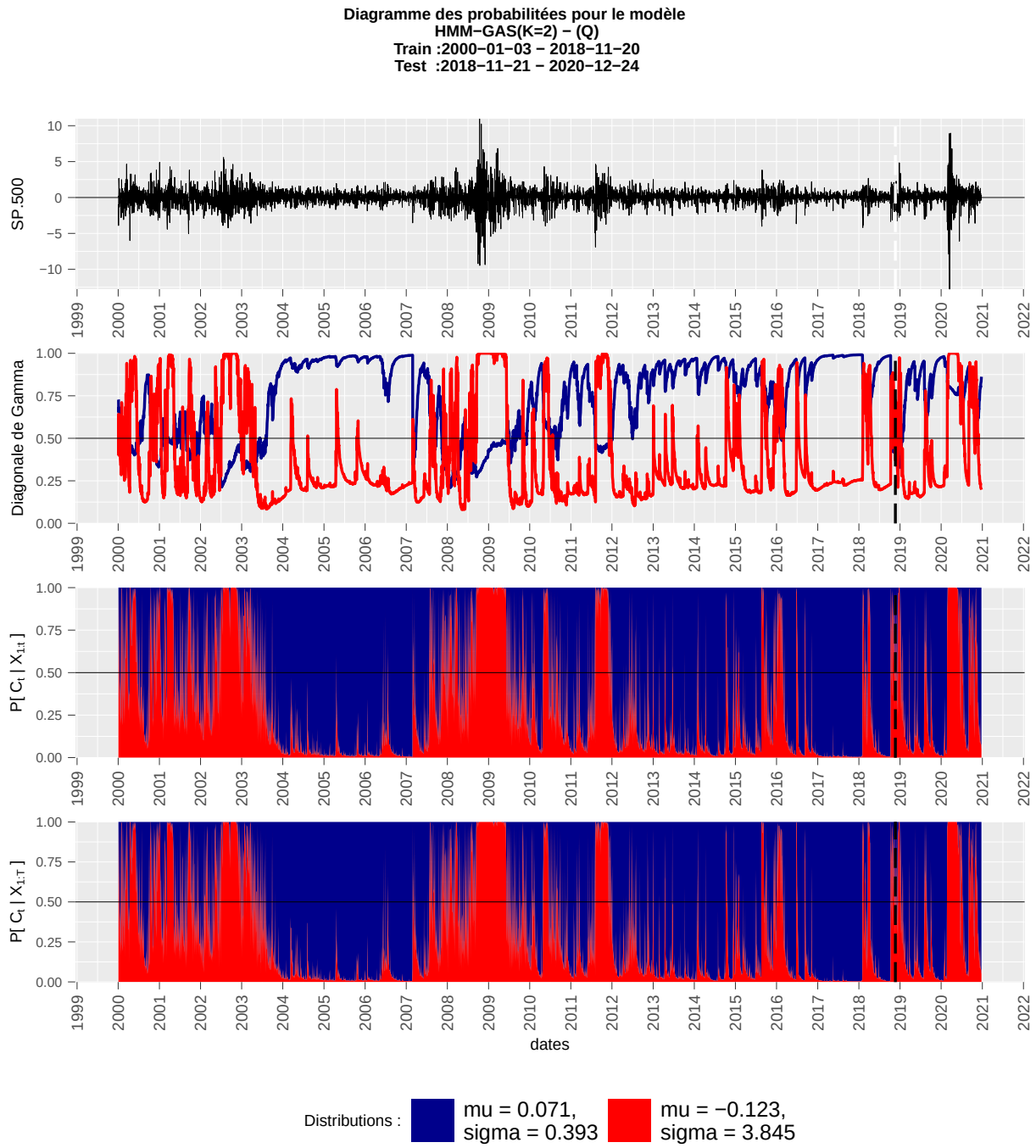


Figure A.8 – Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-GAS($K = 2$) estimées sur le S&P 500

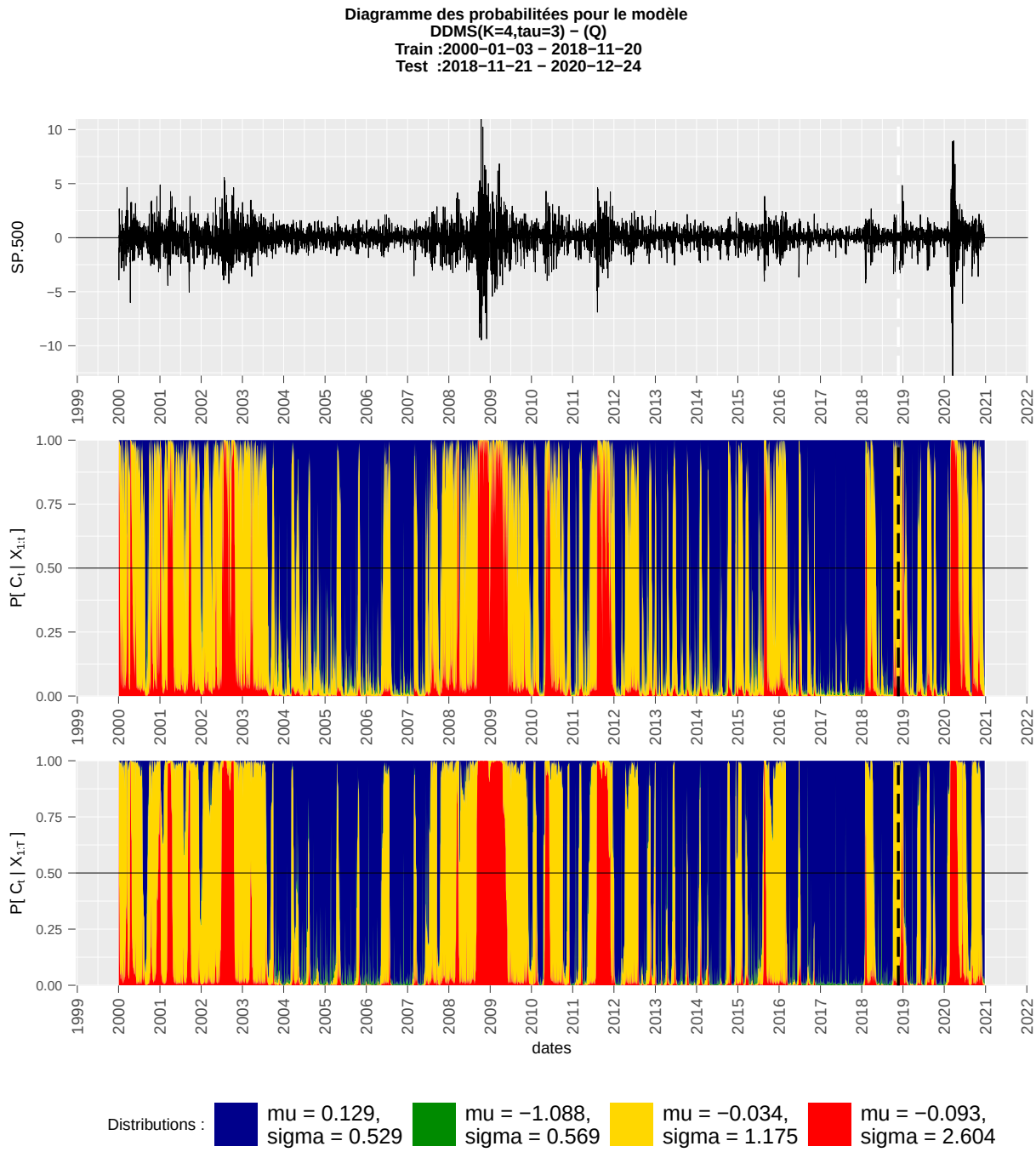


Figure A.9 – Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 3$) estimées sur le S&P500

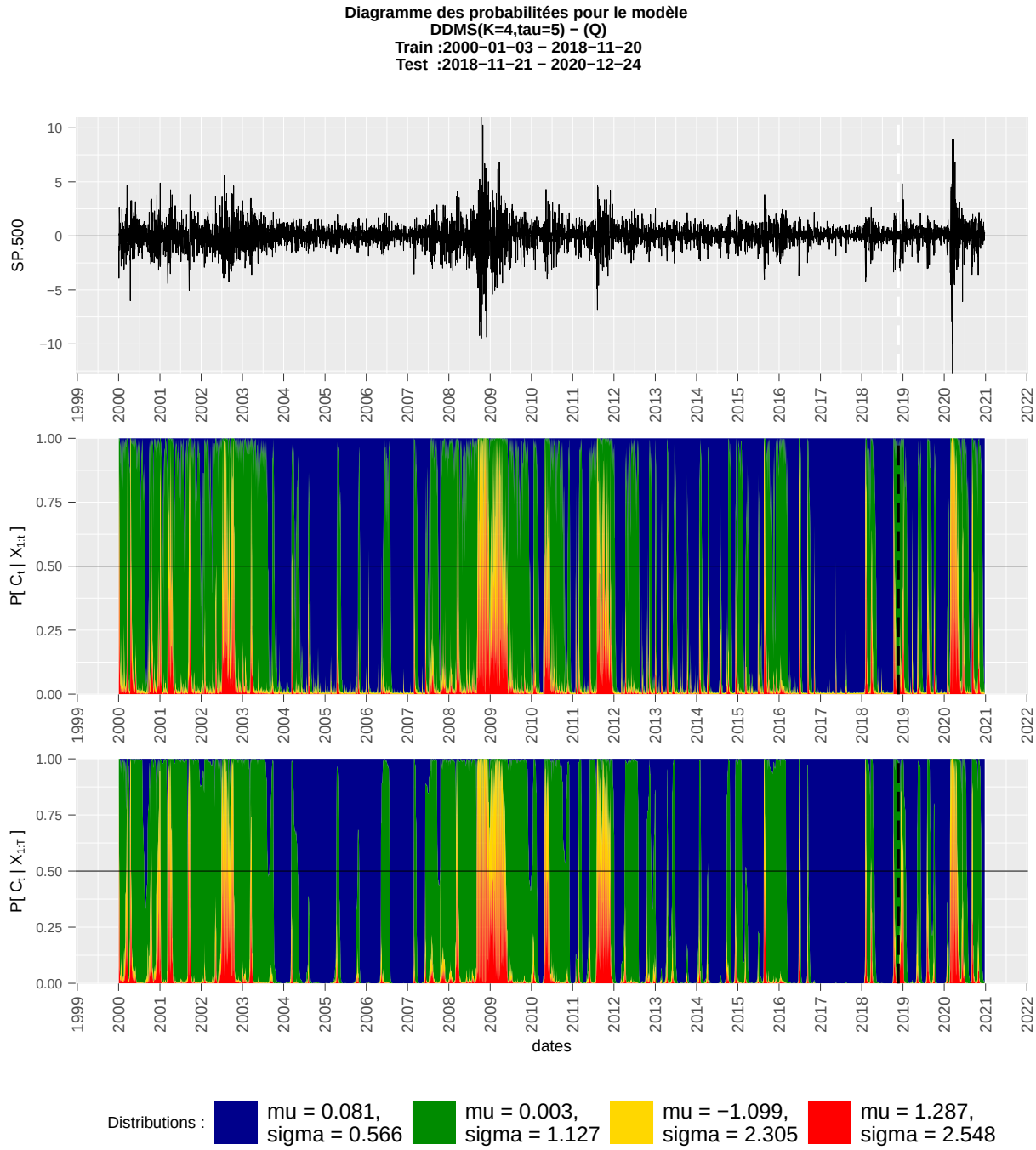


Figure A.10 – Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 5$) estimées sur le S&P500

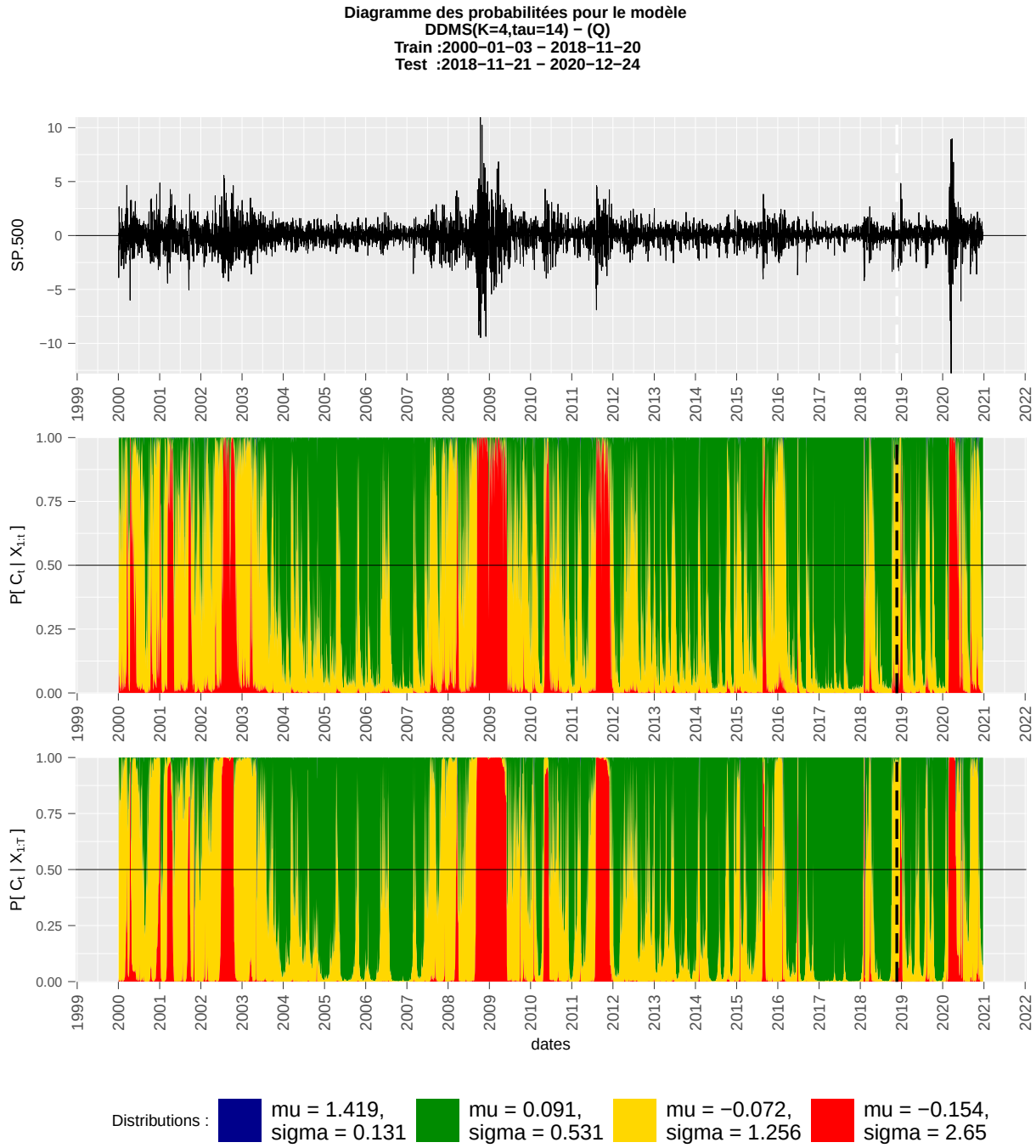


Figure A.11 – Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 14$) estimées sur le S&P 500

A.2. Résultats pour la série du taux de change Canada-États-Unis (CADUSD)

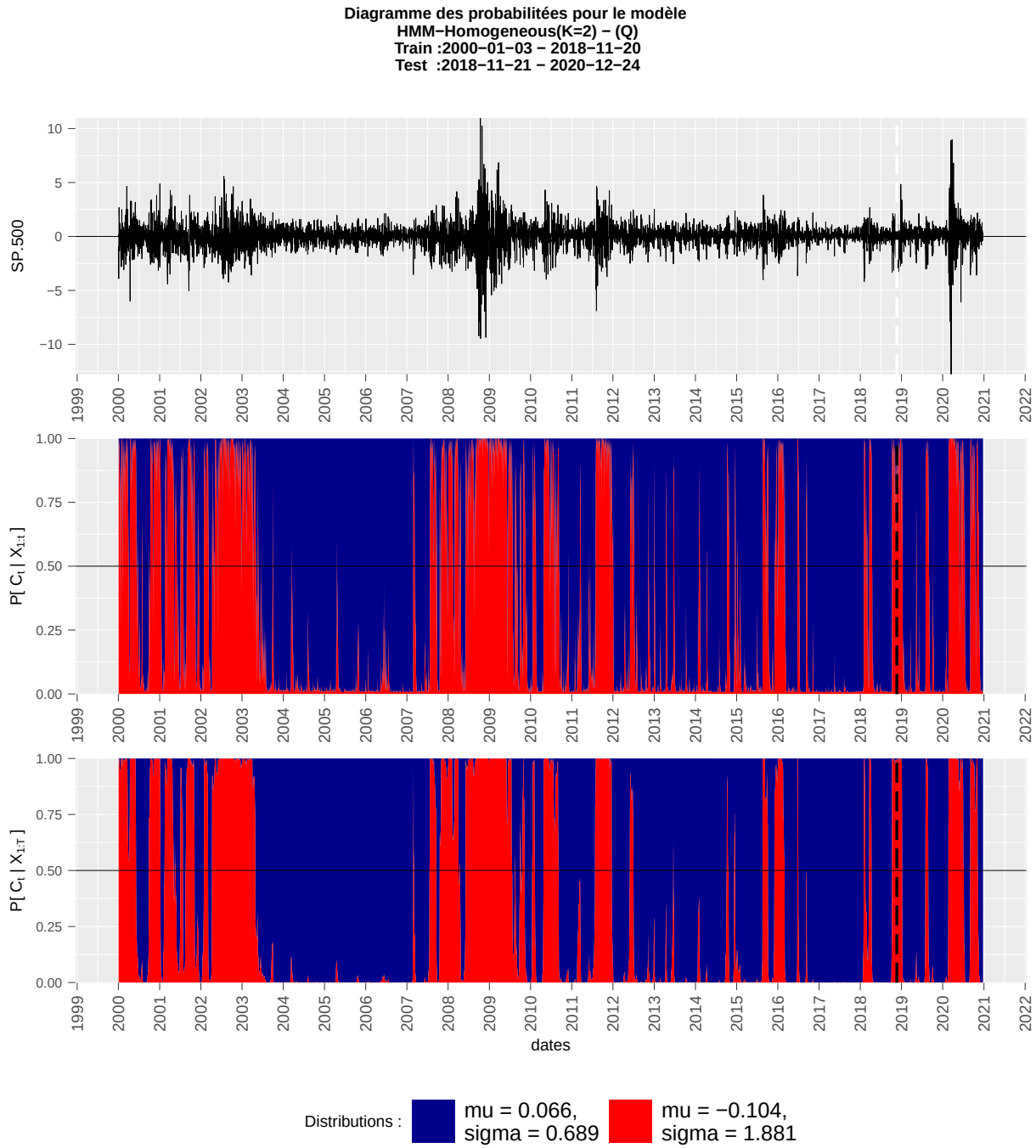


Figure A.12 – Probabilités filtrées et lissées pour le modèle HMM($K = 2$) estimées sur le CADUSD

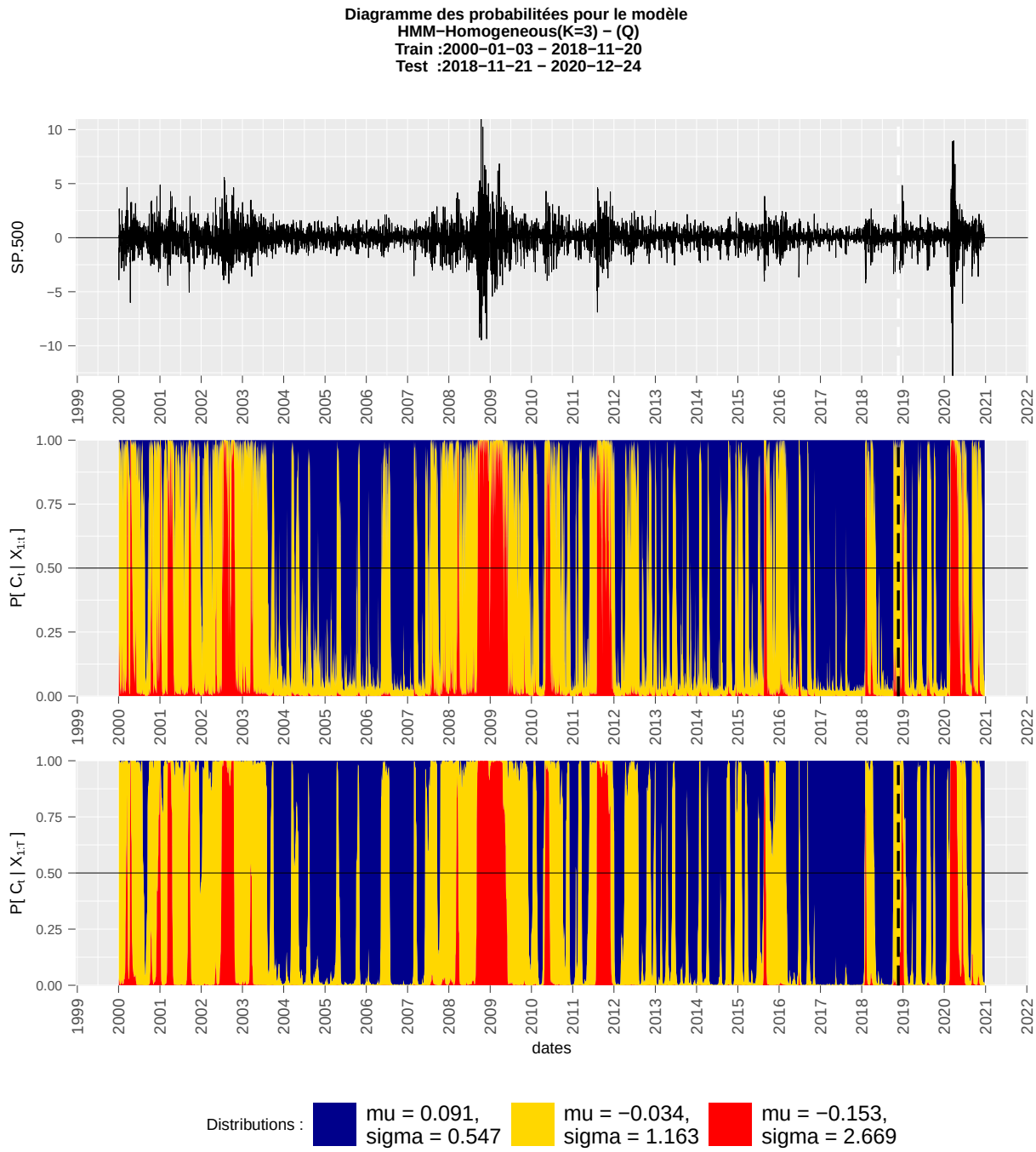


Figure A.13 – Probabilités filtrées et lissées pour le modèle HMM($K = 3$) estimées sur le CADUSD

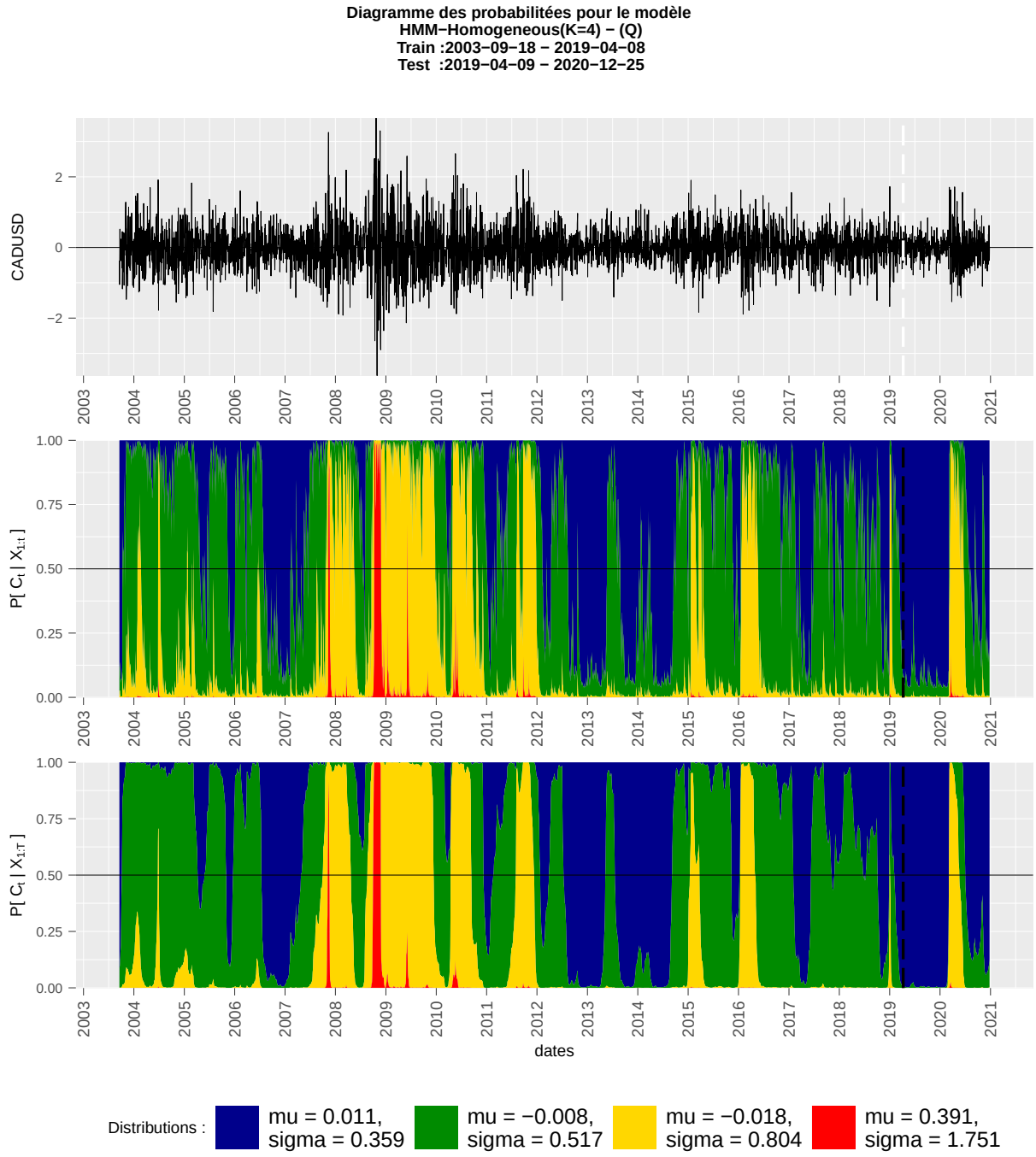


Figure A.14 – Probabilités filtrées et lissées pour le modèle HMM($K = 4$) estimées sur le CADUSD

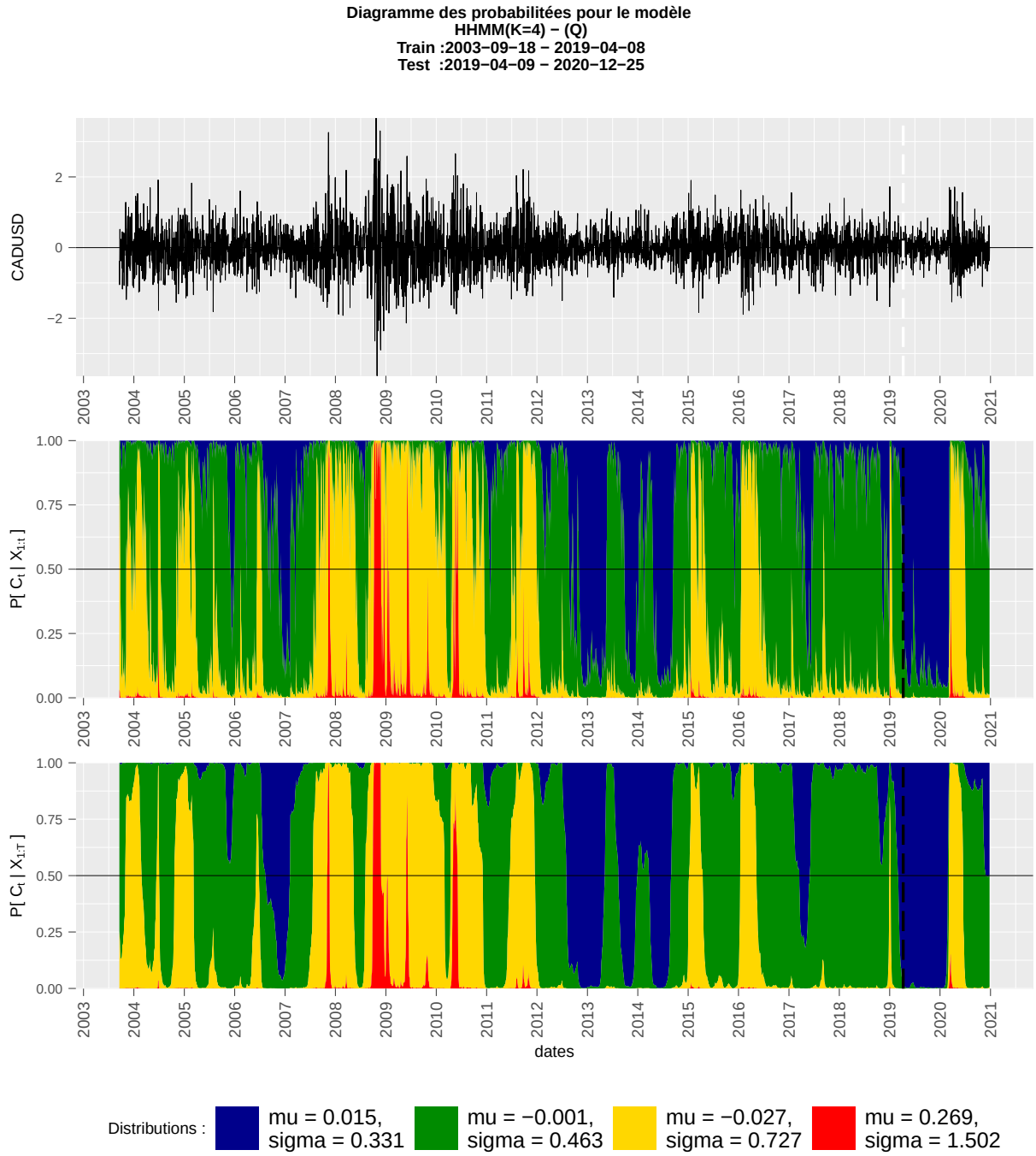


Figure A.15 – Probabilités filtrées et lissées pour le modèle HHMM($K = 4$) estimées sur le CADUSD

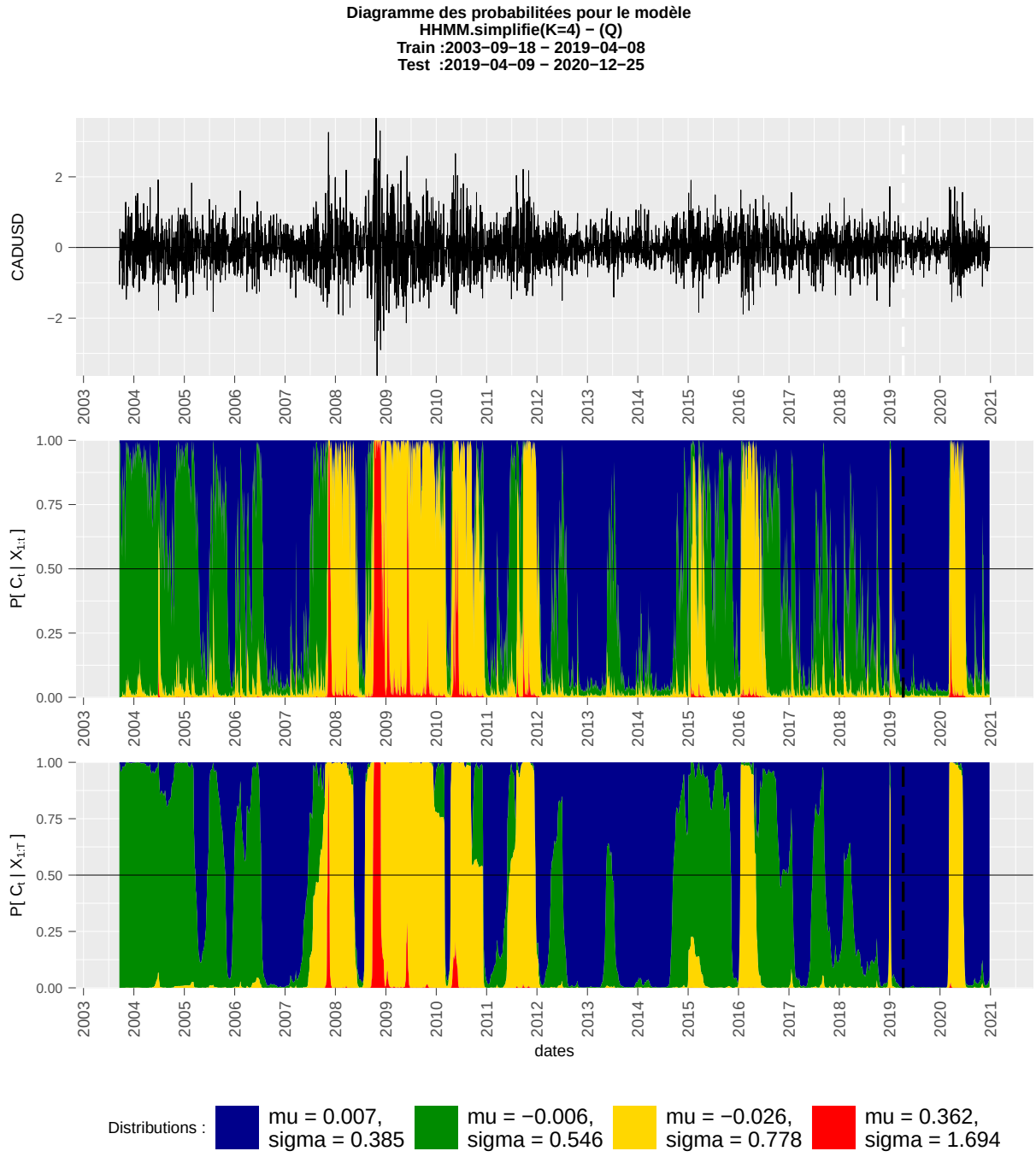


Figure A.16 – Probabilités filtrées et lissées pour le modèle HHMM-simplifié($K = 4$) estimées sur le CADUSD

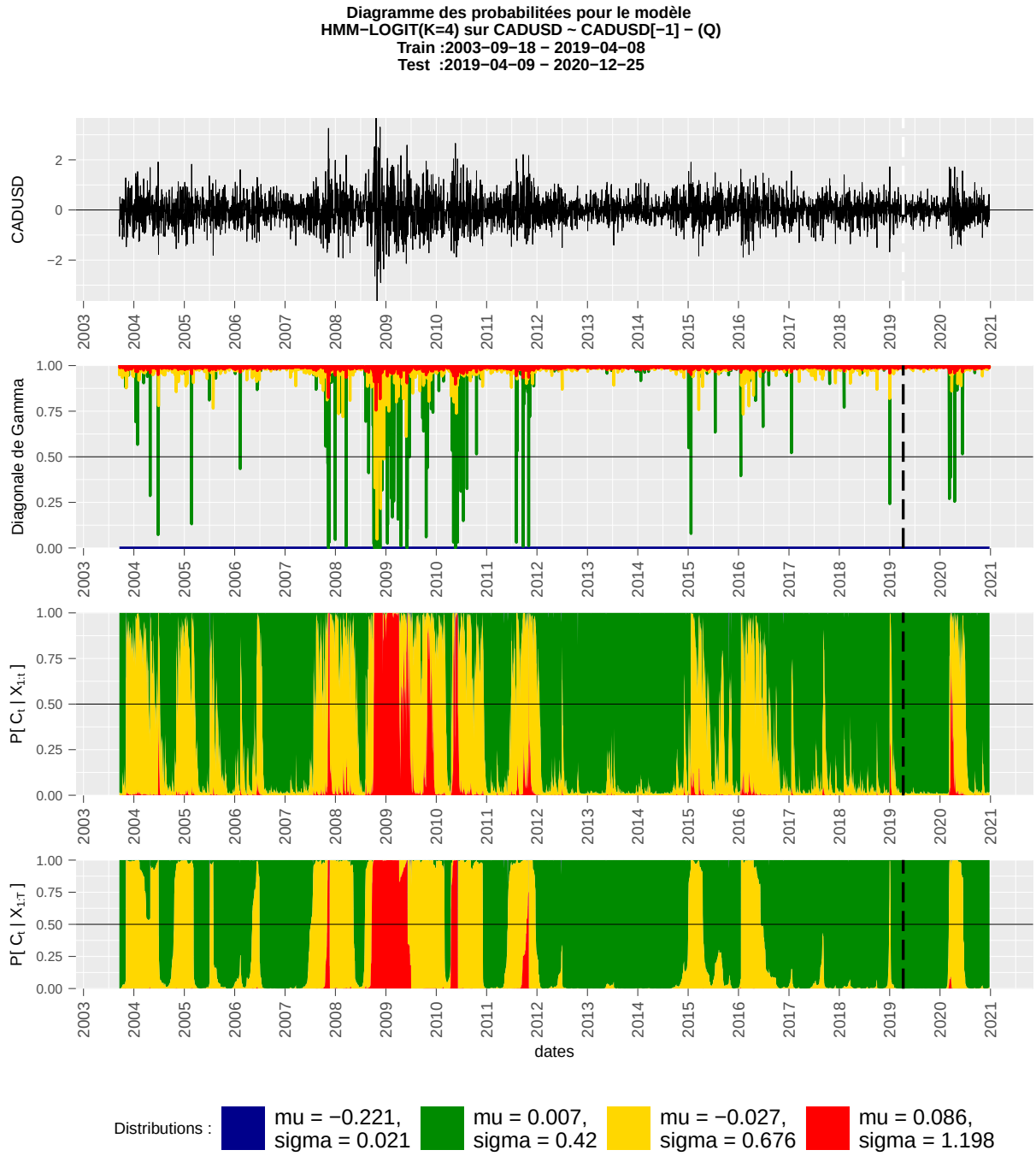


Figure A.17 – Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le CADUSD avec ses valeurs décalées comme variable explicative

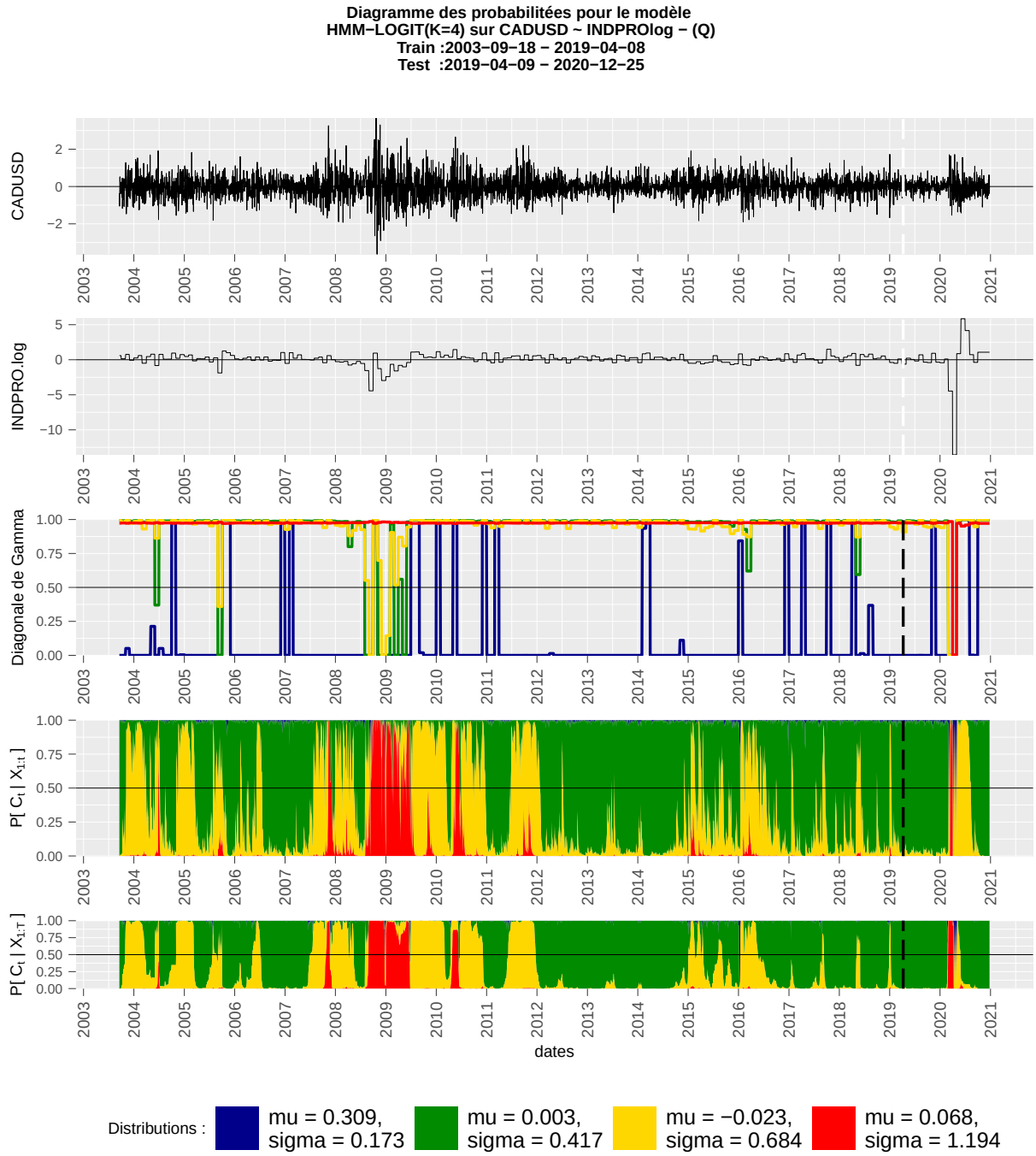


Figure A.18 – Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-LOGIT($K = 4$) estimées sur le CADUSD avec la série INDPRO comme variable explicative

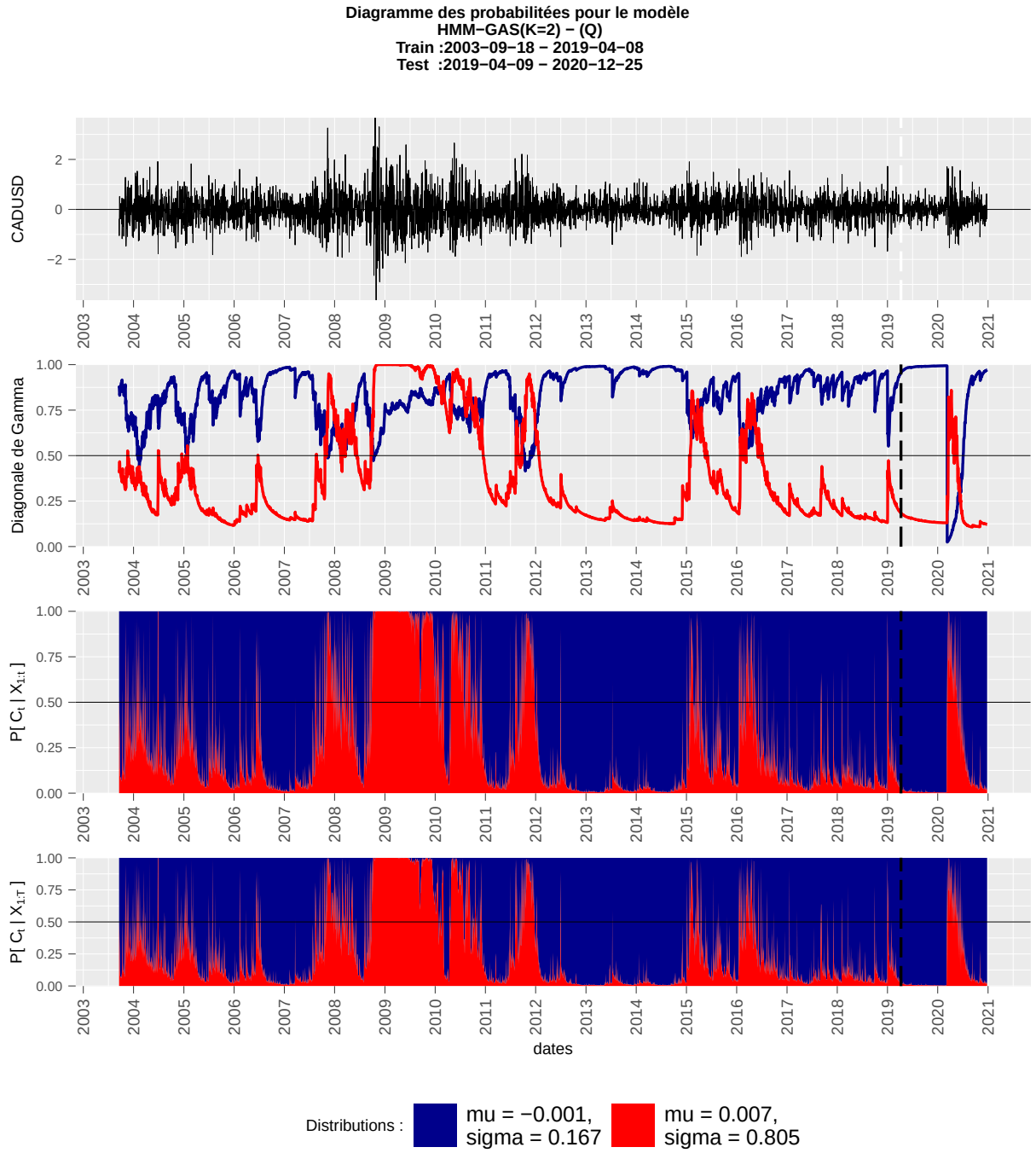


Figure A.19 – Diagonales de $\mathbf{\Gamma}_t$, probabilités filtrées et lissées pour le modèle HMM-GAS($K = 2$) estimées sur le CADUSD

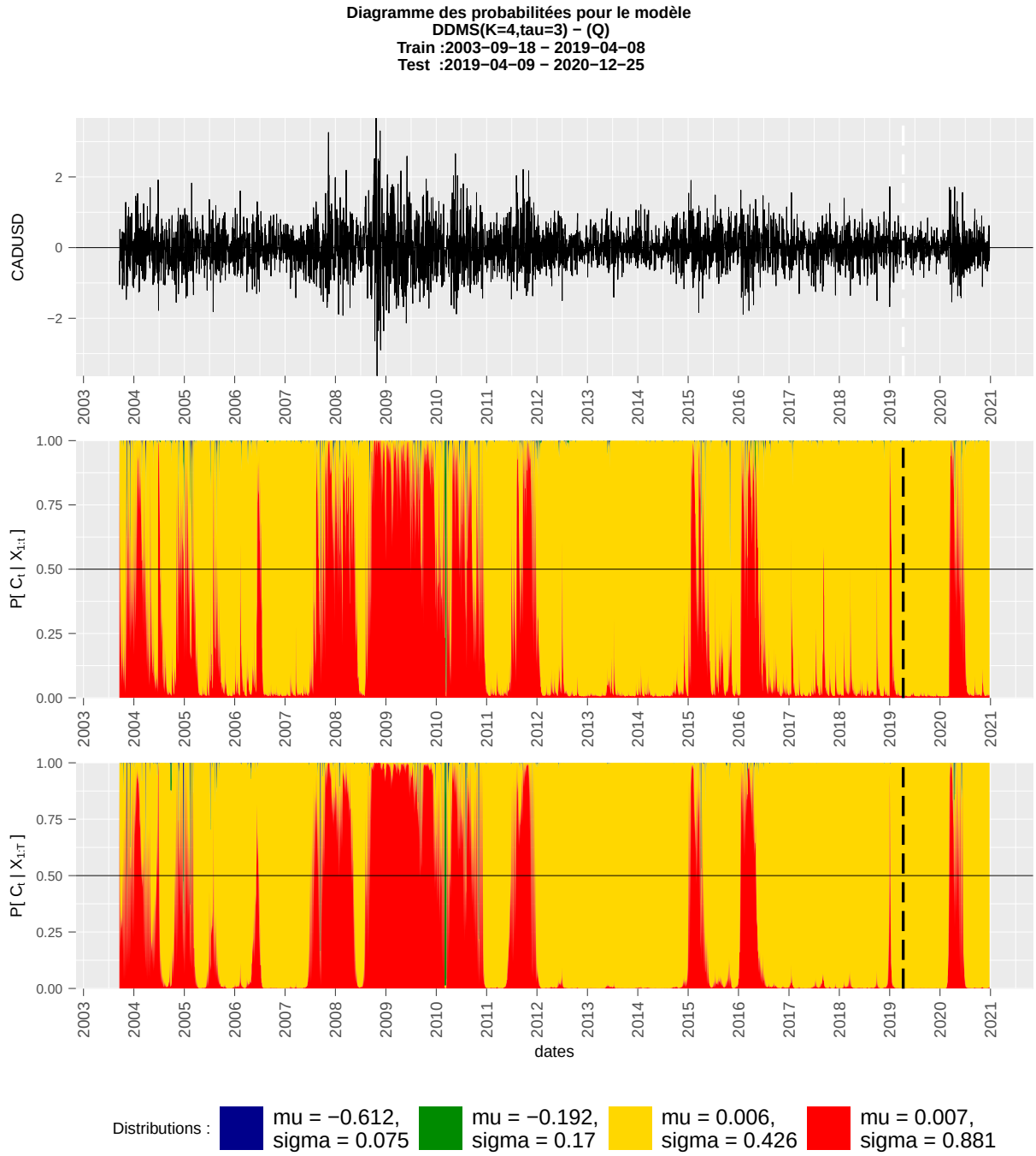


Figure A.20 – Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 3$) estimées sur le CADUSD

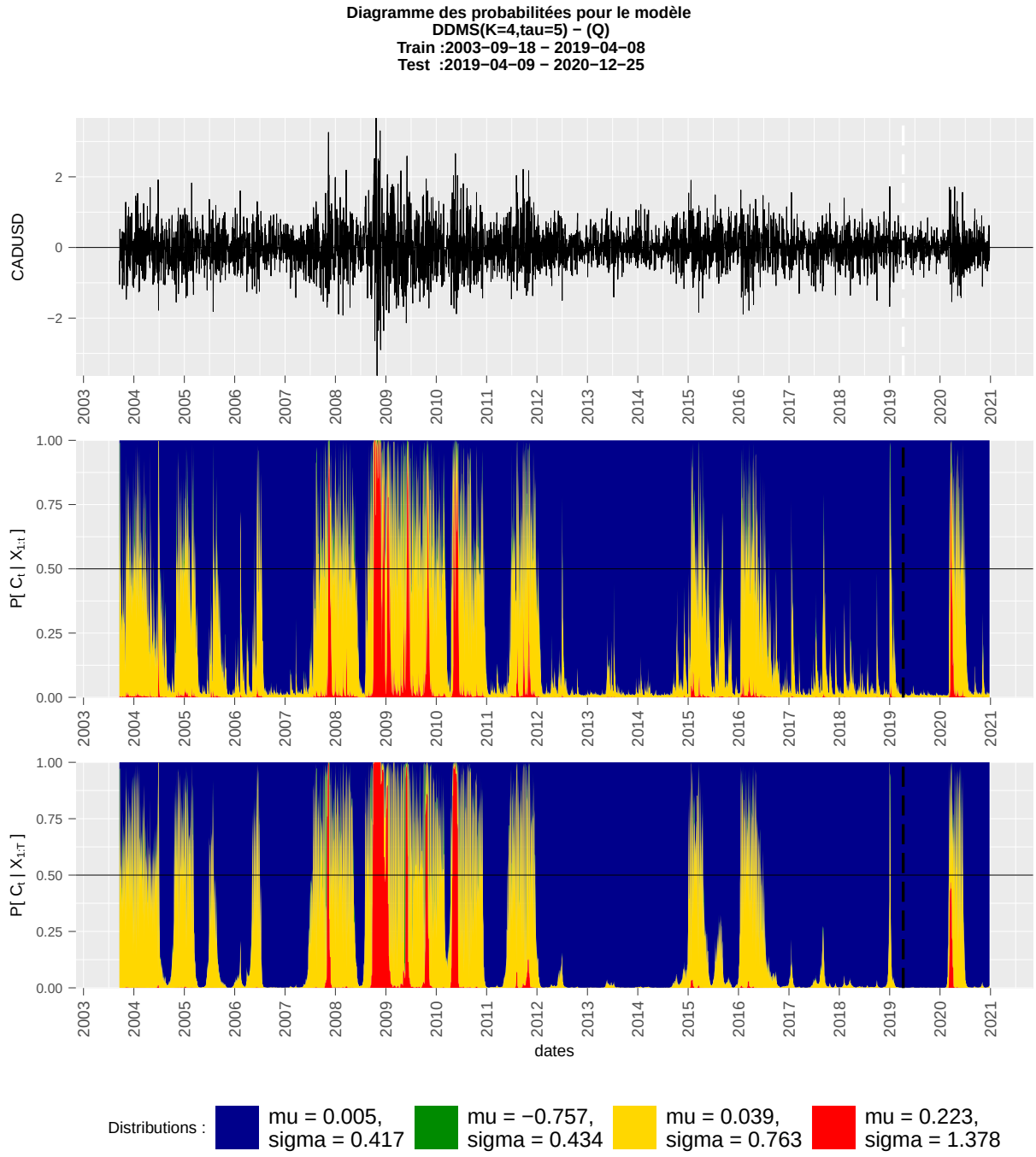


Figure A.21 – Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 5$) estimées sur le CADUSD

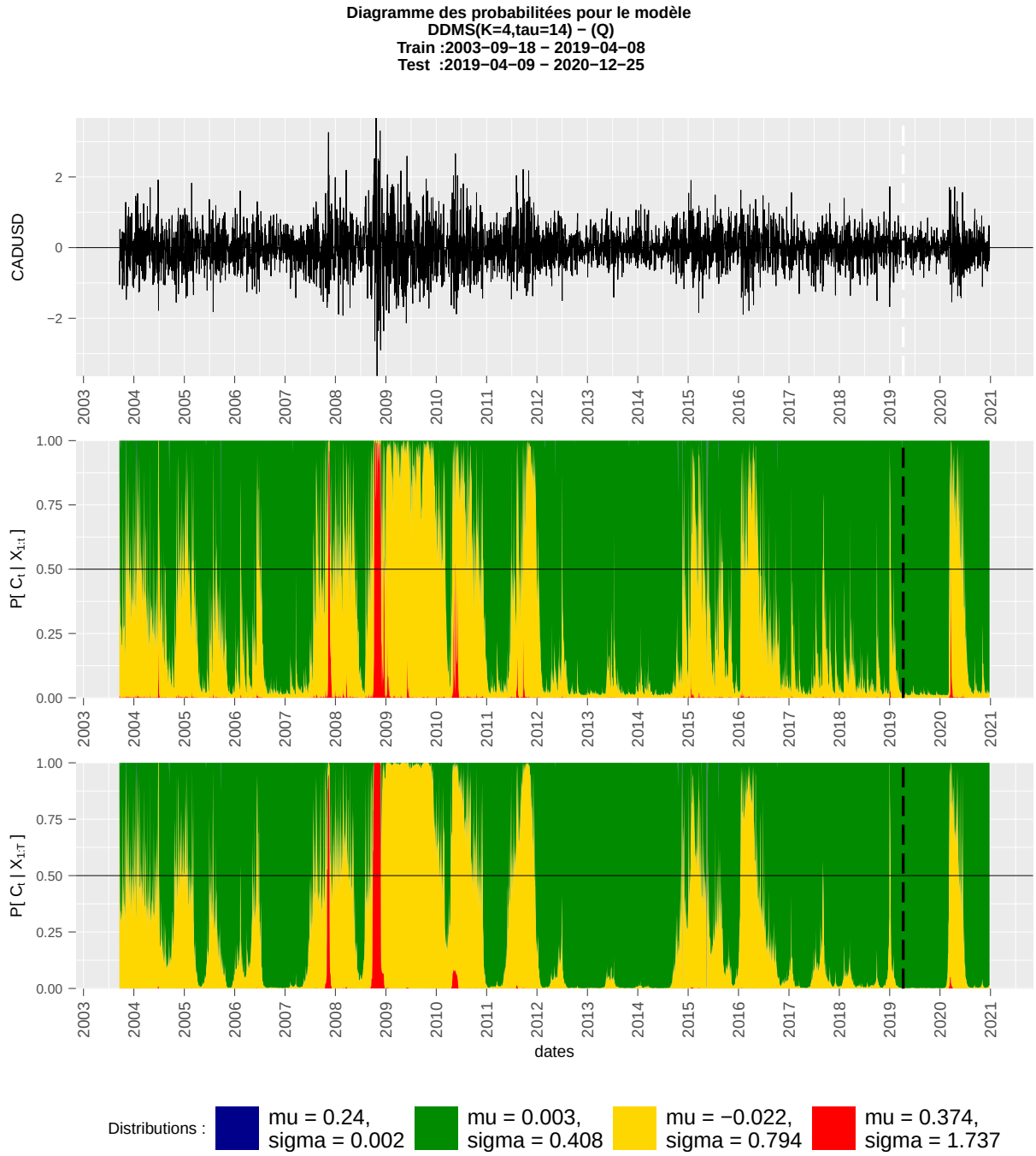


Figure A.22 – Probabilités filtrées et lissées pour le modèle HMM-DDMS($K = 4, \tau = 14$) estimées sur le CADUSD

Université de Montréal

Appendice externe
Modèles de Markov à variables latentes : Matrice de transition non-homogène et reformulation hiérarchique

par

Gabriel Lemyre

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en mathématiques

Orientation mathématiques appliquées

4 janvier 2021

© Gabriel Lemyre, 2021

Appendice I

Propriétés et Résultats supplémentaires concernant les chaînes de Markov

I.1. Propriété des matrices de transition en m pas

Cette section complète l'information présentée à la section 2.2 sur les matrices de transition en m pas. Les probabilités de transition en m pas respectent les conditions suivantes $\forall\, m,t \in \mathbb{N}$:

Propriété 1. $0 \leq \Upsilon_{ij}^{(m)}(t) \leq 1, \forall\, i,j \in \mathcal{C}$,

Propriété 2. $\sum\limits_{j\in\mathcal{C}} \Upsilon_{ij}^{(m)}(t) = 1, \forall\, i \in \mathcal{C}$,

Les propriétés 1 et 2 sont obtenues directement puisque les $\Upsilon_{ij}^{(m)}(t)$ sont des probabilitées et que la somme des probabilités de transition à partir d'un état vers tous les autres doit être égale à 1 par la formule des probabilités totales. Ce sont ces deux propriétés qui font de la matrice $\mathbf{\Gamma}_t^{(m)}$ une matrice stochastique.

Propriété 3 (Équation de Chapman-Kolmogorov)**.**

$$\Upsilon_{ij}^{(m)}(t) = \sum_{h\in\mathcal{C}} \left[\Upsilon_{ih}^{(m-k)}(t) \times \Upsilon_{hj}^{(k)}(t+m-k)\right], \quad \forall\, i,j \in \mathcal{C} \;\; et \; \forall\, k \in \{0,1,\ldots,m\}, \tag{I.1.1}$$

avec

$$\Upsilon_{ij}^{(0)}(t) = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{I.1.2}$$

Il est possible de réécrire cette propriété sous forme matricielle

$$\mathbf{\Gamma}_t^{(m)} = \mathbf{\Gamma}_t^{(m-k)}\mathbf{\Gamma}_{t+m-k}^{(k)}, \quad \forall\, k \in \{0,1,\ldots,m\}, \tag{I.1.3}$$

avec

$$\mathbf{\Gamma}^{(0)} = \mathbf{I}_K. \tag{I.1.4}$$

Démonstration. Sans faire l'hypothèse d'homogénéité de la chaîne de Markov,

$$\begin{aligned}
\gimel_{ij}^{(m)}(t) &= \mathbb{P}[C_{t+m} = j \mid C_t = i], \\
&= \sum_{h\in\mathcal{C}} \mathbb{P}[C_{t+m} = j, C_{t+m-k} = h \mid C_t = i] && \text{par la formule des probabilités totales,} \\
&= \sum_{h\in\mathcal{C}} \mathbb{P}[C_{t+m} = j \mid C_{t+m-k} = h, C_t = i] \\
&\qquad \times \mathbb{P}[C_{t+m-k} = h \mid C_t = i] && \text{par le théorème de Bayes,} \\
&= \sum_{h\in\mathcal{C}} \mathbb{P}[C_{t+m} = j \mid C_{t+m-k} = h] \\
&\qquad \times \mathbb{P}[C_{t+m-k} = h \mid C_t = i] && \text{par la propriété markovienne,} \\
&= \sum_{h\in\mathcal{C}} \left[\gimel_{ih}^{(m-k)}(t) \times \gimel_{hj}^{(k)}(t+m-k)\right], && \forall\, i,j \in \mathcal{C} \text{ et } \forall\, k \in \{0,1,\ldots,m\}.
\end{aligned}$$

□

Lorsque la chaîne de Markov est homogène et par l'équation de Chapman-Kolmogorov, la matrice de transition admet aussi la propriété suivante :

Propriété 4. $\mathbf{\Gamma}^{(m)} = \mathbf{\Gamma}^m$.

Démonstration. En effet, tel que vu à l'équation I.1.3 et en notant que $\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}$, nous pouvons écrire que

$$\begin{aligned}
\mathbf{\Gamma}^{(m)} &= \mathbf{\Gamma}^{(m-1)}\mathbf{\Gamma}, \\
&= \left(\mathbf{\Gamma}^{(m-2)}\mathbf{\Gamma}\right)\mathbf{\Gamma}, \\
&= \ldots \\
&= \mathbf{\Gamma}^m.
\end{aligned}$$

□

I.2. Classification des états

Cette section présente les outils de caractérisation des chaînes de Markov. Ces outils inclus l'accessibilité entre les états d'une chaîne, leurs types de récurrence ainsi que leurs périodes. Ce sont ces caractéristiques qui permettent ensuite de définir l'ergodicité d'une chaîne de Markov.

Définition I.2.1 (Communication entre les états)**.** *Un état j est dit accessible à partir d'un état i, noté $i \longrightarrow j$, si et seulement si $\exists\ m(i,j) \in \mathbb{N}^\star$ t.q. $\mathbb{Y}_{ij}^{(m(i,j))} > 0$. Si, de plus, i est accessible à partir de j, nous dirons que les deux états communiquent et cette relation sera notée $i \longleftrightarrow j$.*

Cette relation, appelée relation d'équivalence, admet la propriété de transitivité telle qu'illustrée à la propriété 3 des matrices de transitions en m pas, présentée en appendice I.1. En effet, si $i \longleftrightarrow k$ et que $k \longleftrightarrow j$, alors on a que $i \longleftrightarrow j$.

Si tous les états de la chaîne communiquent entre-eux, ils sont dits faire partie de la même classe d'équivalence. Lorsque l'espace d'états est composé d'une seule classe d'équivalence, la chaîne est dite ***irréductible****.*

Définition I.2.2 (État récurrent)**.** *Un état k d'une chaîne de Markov est dit* ***récurrent*** *si*

$$f_{kk} \overset{\text{def}}{=} \mathbb{P}[\,\textit{retour à l'état } k \mid \textit{départ de l'état } k\,] = 1. \tag{I.2.1}$$

Sinon ($f_{kk} < 1$), l'état est dit ***transient****.*

Définition I.2.3 (État récurrent positif)**.** *Un état k d'une chaîne de Markov est dit* ***récurrent positif*** *si*

$$r_k \overset{\text{def}}{=} \mathbb{E}[\,\textit{temps de premier retour à l'état } k \mid \textit{départ de l'état } k\,] < \infty. \tag{I.2.2}$$

Sinon ($r_k = \infty$), l'état est dit ***récurrent nul****.*

Définition I.2.4 (Période)**.** *La période d'un état k d'une chaîne de Markov est définie comme*

$$d(k) \overset{\text{def}}{=} \text{pgcd}\Big(n \geq 1 : \mathbb{Y}_{kk}^{(n)} > 0\Big). \tag{I.2.3}$$

Un état k est alors dit ***apériodique*** *si $d(k) \in \{0, 1\}$ et* ***périodique*** *lorsque $d(k) > 1$. Nous noterons $d(k) = 0$ pour indiquer que l'état k n'est pas accessible à partir de lui-même.*

Théorème I.2.1. *Si $i \longleftrightarrow j$, alors*

a) i et j ont la même période,
b) i est récurrent si et seulement si j est récurrent,
c) i est récurrent positif si et seulement si j est récurrent positif.

Si, de plus, l'espace d'états est fini, alors

d) tous les états récurrents sont récurrents positifs,
e) dans tous les cas il y a au moins un état récurrent.

Corollaire I.2.1. *Si l'espace d'états est fini, tous les états d'une chaîne de Markov irréductible sont récurrents positifs.*

Démonstration. Theorème 2 et lemme 7 de Grimmett et Stirzaker [2001], ou alternativement les propositions 1, 2 et 3 ainsi que le corollaire de la section 1.6.4 de Lessard [2014] (démonstrations en 1.8.2, 1.8.3 et 1.8.4). □

Théorème I.2.2. *Soit $i \in \mathcal{C}$, un état d'une chaîne irréductible et d, sa période (et donc celle de tous les états de la chaîne). Alors, $\forall\, j \in \mathcal{C}$, il existe un entier unique $r_j \in [0,d)$ tel que*

a) $\gamma_{ij}^{(s)} > 0 \implies s \equiv r_j (mod\ d)$,
b) $\gamma_{ij}^{(kd+r_j)} > 0, \quad \forall k \geq N(j)$, avec $N(j) \in \mathbb{N}^\star$.

Démonstration. Voir le théorème 1.3 de Seneta [2006]. □

Définition I.2.5 (Ergodicité)**.** *Une chaîne de Markov irréductible, apériodique et sur un espace d'états fini est dite* ***ergodique****.*

I.3. Diagonalisation et le théorème de Perron-Frobenius

Définition I.3.1 (Diagonalisation)**.** *Une matrice carrée $\mathbf{T}$ de dimension $K \times K$ est diagonalisable si elle admet K vecteurs propres linéairement indépendants (ou K valeurs propres distinctes). Elle peut alors s'écrire sous la forme*

$$\mathbf{T} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}, \tag{I.3.1}$$

où la matrice $\mathbf{P}$*, appelée matrice de passage, contient sur ses colonnes lesdits* K *vecteurs propres, la matrice* $\mathbf{\Lambda}$ *est une matrice diagonale dont les coefficients sont les valeurs propres associées à ces vecteurs et* $\mathbf{P}^{-1}$ *est l'inverse de la matrice de passage. Les colonnes de* $\mathbf{P}$ *sont appelées vecteurs propres droits et les lignes de* $\mathbf{P}^{-1}$ *sont appelées vecteurs propres gauches.*

Afin d'alléger l'écriture, nous adopterons la notation suivante,

- $\lambda(\mathbf{T})$ *: L'ensemble des valeurs propres de* $\mathbf{T}$*, aussi appelé le spectre de* $\mathbf{T}$*,*
- λ_i *: Le* i^e *élément de* $\lambda(\mathbf{T})$*,*
- $\boldsymbol{\nu}_i$ *: Le vecteur propre droit associé à* λ_i*,*
- $\vec{\omega}_i$ *: Le vecteur propre gauche associé à* λ_i*.*

Les valeurs propres sont obtenues en trouvant les valeurs de λ *telles que*

$$\det\left(\mathbf{T} - \lambda \mathbf{I}_K\right) = 0. \tag{I.3.2}$$

Les vecteurs propres sont obtenus en trouvant les vecteurs $\boldsymbol{\nu}$ *tels que* $\mathbf{T}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$*, ce qui est équivalent à résoudre*

$$(\mathbf{T} - \lambda)\boldsymbol{\nu} = 0. \tag{I.3.3}$$

Prenons en exemple la matrice de transition d'une chaîne de Markov et calculons la puissance 2 de la matrice $\mathbf{\Gamma}$,

$$\begin{aligned}
\mathbf{\Gamma}^2 = \mathbf{\Gamma}\mathbf{\Gamma} &= \Big(\mathbf{P}\Lambda\mathbf{P}^{-1}\Big)\Big(\mathbf{P}\Lambda\mathbf{P}^{-1}\Big), \\
&= \mathbf{P}\Lambda\Big(\mathbf{P}^{-1}\mathbf{P}\Big)\Lambda\mathbf{P}^{-1}, \\
&= \mathbf{P}\Lambda\mathbf{I}_K\Lambda\mathbf{P}^{-1}, \\
&= \mathbf{P}\Lambda^2\mathbf{P}^{-1}, \text{puisque } \Lambda \text{ est diagonale.}
\end{aligned}$$

Selon cette même logique, nous pouvons écrire en général que

$$\mathbf{\Gamma}^m = \mathbf{P}\Lambda^m\mathbf{P}^{-1}. \tag{I.3.4}$$

Cette diagonalisation permet donc d'écrire que

$$\begin{aligned}\lim_{m\to\infty}\boldsymbol{\Gamma}^{(m)} &= \lim_{m\to\infty}\mathbf{P}\Lambda^m\mathbf{P}^{-1},\\ &= \mathbf{P}\lim_{m\to\infty}\Lambda^m\mathbf{P}^{-1},\\ &= \mathbf{P}\lim_{m\to\infty}\begin{bmatrix}\lambda_1^m & 0 & \cdots & 0\\ 0 & \lambda_2^m & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_K^m\end{bmatrix}\mathbf{P}^{-1}. \qquad \text{(I.3.5)}\end{aligned}$$

Une simplification supplémentaire et très importante nous provient du théorème de Perron-Frobenius et d'un de ses corollaires, présentés entre autres dans Seneta [2006]. Mais d'abord,

Définition I.3.2 (Matrice *primitive*)**.** *Une matrice non-négative* $\mathbf{T}$ *est dites primitive si* $\exists k \in \mathbb{N}$ *tel que* $\mathbf{T}^k > 0$.

Théorème I.3.1. *Une matrice de transition associée à une chaîne de Markov irréductible et apériodique est primitive et inversement.*

Démonstration. Si la matrice de transition $\boldsymbol{\Gamma}$ est irréductible avec période $d = 1$, alors par le théorème I.2.2, nous avons que

$$\gamma_{ij}^{(k)} > 0,\ \forall k \geq N(i,j).$$

Ainsi, en prenant $N^* = \max_{i,j} N(i,j)$, nous avons que

$$\gamma_{ij}^{(k)} > 0,\ k \geq N^*,\ \forall\, i,j \in \mathcal{C}.$$

Ce qui revient à dire que

$$\boldsymbol{\Gamma}^k > 0, \forall k \geq N^*$$

et que donc, la matrice $\boldsymbol{\Gamma}$ est primitive.

Si maintenant la matrice $\boldsymbol{\Gamma}$ est primitive, nous avons directement que la matrice $\boldsymbol{\Gamma}$ est irréductible par définition. De plus, nous avons qu'il $\exists m \in \mathbb{N}$ tel que $\boldsymbol{\Gamma}^m > 0$. Supposons que $\boldsymbol{\Gamma}^{m+1}$ ne soit pas positive. Si $\boldsymbol{\Gamma}^{m+1}$ n'est pas positive, $\exists i,j \in \mathcal{C} \mid \gamma_{ij}^{(m+1)} = 0$. Puisque

toutes les entrées de $\mathbf{\Gamma}^m$ sont positives et que $\mathbf{\Gamma} \geq 0$, la seule manière qu'une entrée de $\mathbf{\Gamma}^{m+1}$ soit nulle est qu'une colonne complète de $\mathbf{\Gamma}$ soit nulle. Ceci signifierait qu'un état de la chaîne n'est accessible depuis aucun autre état, ce qui est en contradiction avec le fait que la chaîne associée à $\mathbf{\Gamma}$ est irréductible.

Ainsi, nous avons que $\mathbf{\Gamma}^{m+1} > 0$ et puisque le pgcd$(m,m+1) = 1$, la chaîne associée est apériodique. □

Le théorème suivant s'applique donc dans le cas de matrices de transition associées à des chaînes de Markov irréductibles et apériodiques sur un espace d'états fini.

Théorème I.3.2 (Perron-Frobenius)**.** *Soit* $\mathbf{T}$*, une matrice non-négative, primitive et de dimension* $K \times K$*. Elle admet alors une valeur propre* r *telle que*

(a) $r \in \mathbb{R}_+^*$,
(b) $r > |\lambda|, \quad \forall\, \lambda \neq r \in \lambda(\mathbf{T})$,
(c) r *est une racine de multiplicité algébrique un du polynôme caractéristique de* $\mathbf{T}$ (*i.e. solution à* $\det(\lambda \mathbf{I}_K - \mathbf{T}) = 0$),
(d) Si $0 \leq \mathbf{B} \leq \mathbf{T}$ *et que* β *est une valeur propre de* $\mathbf{B}$*, alors* $|\beta| \leq r$*. De plus, si* $|\beta| = r$*, alors certainement* $\mathbf{B} = \mathbf{T}$,
(e) On peut associer à r *des vecteurs propres droit et gauche strictement positifs,*
(f) Les vecteurs propres associés à r *sont uniques à une constante multiplicative près, ainsi* r *a une multiplicité géométrique de un (* *i.e.* $\nexists$ *d'autre vecteur linéairement indépendant de* $\boldsymbol{\nu}_1$ *et* $\vec{\omega}_1$ *associé à la valeur propre* λ_1*).*

Corollaire I.3.1.

$$\min_j \sum_{j=1}^K t_{ij} \leq r \leq \max_i \sum_{j=1}^K t_{ij}, \tag{I.3.6}$$

avec l'égalité des deux côtés seulement si les sommes sur les lignes de la matrice $\mathbf{T}$ *sont égales.*

Les démonstrations du théorème I.3.2 ainsi que du corollaire I.3.1 sont très longues et seront donc omises dans cet ouvrage. Une approche algébrique peut être retrouvée à la section 1.1 de Seneta [2006].

Nous nous limiterons au cas particulier où la chaîne est ergodique (période $d = 1$) et noterons pour la suite $\lambda_1 := r$.

Théorème I.3.3. *Soit $\mathbf{\Gamma}$, la matrice de transition d'une chaîne de Markov ergodique. Alors,*

(a) $\lambda_1 = 1$,
(b) $\lambda_1 > |\lambda_i|, \quad \forall$ *valeur propre* λ_i *avec* $i \in \{2, \ldots, K\}$,
(c) λ_1 *est une racine de multiplicité algébrique un du polynôme caractéristique de* $\mathbf{\Gamma}$ (*i.e. solution à* $\det(\lambda \mathbf{I}_K - \mathbf{\Gamma}) = 0$),
(d) On peut associer à λ_1 *des vecteurs propres droit* $\boldsymbol{\nu}_1$ *et gauche* $\vec{\omega}_1$ *strictement positifs,*
(e) Les vecteurs propres associés à λ_1 *sont uniques à une constante multiplicative près, ainsi* λ_1 *a une multiplicité géométrique de un.*

Démonstration. Ce théorème est une conséquence directe du théorème de Perron-Frobenius ainsi que du corollaire I.3.1. □

Nous avons donc que $\lambda_1 = 1$ et $|\lambda_i| < \lambda_1, \forall i \neq 1$. Puisque nous savons que la vitesse de convergence de la matrice $\mathbf{\Gamma}$ dépend seulement de ses valeurs propres, tel qu'illustrée à l'équation (I.3.5), il convient de s'intéresser à la valeur prise par la deuxième plus grande valeur propre en valeur absolue.

Théorème I.3.4. *Soit m_i, la multiplicité algébrique de λ_i.*
Pour une matrice primitive $\mathbf{\Gamma}$:

(a) Si $\lambda_2 \neq 0$, alors lorsque $k \to \infty$

$$\mathbf{\Gamma}^k = r^k \boldsymbol{\nu}_1 \vec{\omega}_1 + \mathrm{O}\Big(k^{m_2 - 1} |\lambda_2|^k\Big), \tag{I.3.7}$$

$$= \mathbf{1}_K \vec{\pi} + \mathrm{O}\Big(k^{m_2 - 1} |\lambda_2|^k\Big). \tag{I.3.8}$$

(a) Si $\lambda_2 = 0$, alors

$$\mathbf{\Gamma}^k = r^k \boldsymbol{\nu}_1 \vec{\omega}_1, \tag{I.3.9}$$

$$= \mathbf{1}_K \vec{\pi}. \tag{I.3.10}$$

Ainsi, peu importe la valeur prise par λ_2, nous avons que

$$\lim_{k \to \infty} \mathbf{\Gamma}^k \to \mathbf{1}_K \vec{\pi}, \tag{I.3.11}$$

puisque $\lambda_1 \overset{\text{def}}{=} r = 1$.

Démonstration. Ce théorème est obtenu en combinant les théorèmes 1.2 et 4.2 de Seneta [2006]. □

Remarque I.3.1. Le théorème de Perron-Frobenius est très important car c'est grâce à lui que l'on peut garantir l'existence et l'unicité de la distribution stationnaire d'une chaîne de Markov ergodique. Le corrolaire I.3.1 nous permet de garantir que la distribution stationnaire de la chaîne correspond au vecteur propre gauche normalisé associé à la valeur propre $\lambda_1 = 1$.

Nous pouvons donc écrire la matrice de transition d'une chaîne de Markov ergodique sous la forme

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{1}_K & \boldsymbol{\nu}_2 & \cdots & \boldsymbol{\nu}_K \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_K \end{bmatrix} \begin{bmatrix} \vec{\pi} \\ \vec{\omega}_2 \\ \cdots \\ \vec{\omega}_K \end{bmatrix} \tag{I.3.12}$$

et l'équation (I.3.5) devient

$$\begin{aligned} \lim_{m\to\infty} \boldsymbol{\Gamma}^{(m)} &= \begin{bmatrix} \mathbf{1}_K & \boldsymbol{\nu}_2 & \cdots & \boldsymbol{\nu}_K \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \vec{\pi} \\ \vec{\omega}_2 \\ \cdots \\ \vec{\omega}_K \end{bmatrix} \\ &= \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_K \\ \pi_1 & \pi_2 & \cdots & \pi_K \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_K \end{bmatrix}. \end{aligned}$$

Une version de ce théorème pour les chaînes de Markov irréductibles sur un espace d'états fini est présentée par Grimmett et Stirzaker [2001]. Ainsi, les affirmations de la remarque I.3.1 s'appliquent aussi lorsque la chaîne est périodique.

I.4. Distribution initiale et conjointe

Une chaîne de Markov est complètement déterminée par sa matrice de transition ainsi que la distribution initiale de la chaîne, soit celle de C_1. En effet, il est possible d'obtenir

toutes les distributions marginales et conjointes de la chaîne grâce à la propriété markovienne et aux probabilités de transition conditionnelles. En conditionnant sur le premier état de la chaîne et décomposant les probabilités de transition un pas à la fois, nous obtenons par récurrence que la fonction de masse conjointe des variables $C_1, \ldots, C_n$ est simplement

$$\begin{aligned}
&\mathbb{P}[C_n = i_n, \ldots, C_1 = i_1] \\
&\quad = \mathbb{P}[C_n = i_n \mid C_{n-1} = i_{n-1}, \ldots, C_1 = i_1] \times \mathbb{P}[C_{n-1} = i_{n-1}, \ldots, C_1 = i_1], \\
&\quad = \mathbb{P}[C_n = i_n \mid C_{n-1} = i_{n-1}] \times \mathbb{P}[C_{n-1} = i_{n-1}, \ldots, C_1 = i_1], \\
&\quad = \gamma_{i_{n-1},i_n} \times \mathbb{P}[C_{n-1} = i_{n-1}, \ldots, C_1 = i_1], \\
&\quad = \gamma_{i_{n-1},i_n} \times \gamma_{i_{n-2},i_{n-1}} \times \mathbb{P}[C_{n-2} = i_{n-2}, \ldots, C_1 = i_1], \\
&\quad = \gamma_{i_{n-1},i_n} \times \gamma_{i_{n-2},i_{n-1}} \times \cdots \times \gamma_{i_1,i_2} \times \mathbb{P}[C_1 = i_1], \quad \forall\, i_1, \ldots, i_n \in \mathcal{C},
\end{aligned}$$

où les probabilités de transition γ_{i_{n-1},i_n} seraient simplement notées $\gamma_{i_{n-1},i_n}(n-1)$ lorsque la chaîne n'est pas homogène.

Il est évident que puisque les probabilités conditionnelles sont entre 0 et 1, plus la chaîne est longue et plus la probabilité conjointe tends vers 0. Cet enjeux est discuté à la section 3.5.2, où il est question de méthodes de filtrage permettant que la vraisemblance des réalisations de la chaîne ne soit pas numériquement nulle.

Les distributions marginales des autres pas de la chaîne étant aussi obtenues à partir de la distribution initiale de la chaîne, il est important de faire une hypothèse par rapport à cette initialisation lorsque nous construirons des modèles faisant usage de chaîne de Markov. Ceci est évident pour le filtre d'Hamilton présenté à la section 3.6.1.1. Il est d'abord possible de voir ces probabilités comme $(K-1)$ paramètres du modèles, en obtenant la dernière probabilité par soustraction

$$\vec{\mathfrak{u}}(1) = \left(\mathfrak{u}_1(1), \ldots, \mathfrak{u}_{K-1}(1), 1 - \sum_{k=1}^{(K-1)} \mathfrak{u}_k(1) \right). \tag{I.4.1}$$

Dans le cas où la chaîne de Markov est ergodique, il est aussi possible de choisir la distribution stationnaire comme distribution initiale. Ainsi, nous ne considérons pas celles-ci comme un vecteur de paramètres, ce qui allège le modèle de $(K-1)$ paramètres. Ainsi, nous

noterons dans ce cas

$$\vec{\mathfrak{u}}(1) = \vec{\pi}. \tag{I.4.2}$$

Puisqu'une chaîne ergodique admet une matrice de transition homogène, ce choix signifie que la distribution marginale de la chaîne est constante. En effet,

$$\begin{aligned}\vec{\mathfrak{u}}(t) &= \vec{\mathfrak{u}}(1)\mathbf{\Gamma}^t,\\ &= \vec{\pi}\mathbf{\Gamma}^t,\\ &= \vec{\pi}.\end{aligned}$$

Il y a naturellement plusieurs autres choix possibles. Choisir d'assigner une probabilité égale à chaque état est une bonne manière de ne pas prendre de décision par rapport à l'initialisation de la chaîne. L'inverse fonctionne aussi, c'est-à-dire de simplement choisir aléatoirement un état comme étant le premier et de représenter le vecteur de distribution initiale sous forme canonique, ou comme un vecteur *one-hot*[1]. Par exemple, si nous faisions l'hypothèse que l'état initial de la chaîne est le troisième, nous initialiserions celle-ci par

$$\vec{\mathfrak{u}}(1) = (0,0,1,0,0,\ldots,0). \tag{I.4.3}$$

Peu importe le choix qui sera fait concernant cette distribution, si la chaîne admet une distribution stationnaire, le théorème de Perron-Frobenius (I.3.2) implique que la distribution marginale de la chaîne de Markov converge à vitesse géométrique vers la distribution stationnaire. La vitesse de cette convergence dépend de la valeur de la deuxième plus grande valeur propre de la matrice $\mathbf{\Gamma}$.

I.5. Temps de séjour dans un état

Sous l'hypothèse d'homogénéité, le temps de séjour dans un état de la chaîne est de distribution géométrique. En effet, nous avons que

$$\mathbb{P}[C_{t+1} = i,\ldots,C_{t+k} = i, C_{t+k+1} \neq i \mid C_t = i] = \gamma_{ii}^k(1-\gamma_{ii}). \tag{I.5.1}$$

Ceci signifie que, si la chaîne est dans l'état i au temps t, alors l'espérance du temps de séjour dans cet état à partir de—et en incluant—t est de $1/(1-\gamma_{ii})$.

1. Cette appellation est très utilisée en apprentissage automatique (*Machine Learning*), en particulier lorsqu'il est question de classification, afin de définir la cible d'un modèle cherchant à classer des objets dans des catégories.

I.6. Moments et autocovariance des variables aléatoires C_t

Afin de généraliser les résultats de cette section à des cas où les états de la chaîne ne sont pas simplement les entiers de 1 à K, supposons que l'espace d'états est $\mathcal{C} = \{\upsilon_1, \ldots, \upsilon_K\} \in \mathbb{R}^K$. À cet effet, nous adopterons la notation $\boldsymbol{\upsilon} \stackrel{\text{def}}{=} (\upsilon_1, \ldots, \upsilon_K)'$ et la matrice diagonale $\boldsymbol{\Upsilon} \stackrel{\text{def}}{=} \text{diag}(\boldsymbol{\upsilon})$ afin d'alléger la notation.

I.6.1. Moments de C_t

Les moments d'une chaîne de Markov pour $t \in \mathbb{N}$ sont donnés par

$$\begin{aligned}\mathbb{E}[f(C_t)] &= \sum_{i=1}^{K} f(\upsilon_i)\mathbb{P}[C_t = \upsilon_i], \\ &= \sum_{i=1}^{K} f(\upsilon_i)\mathfrak{u}_i(t), \\ &= \vec{\mathfrak{u}}(t) f(\boldsymbol{\upsilon}), \end{aligned} \tag{I.6.1}$$

avec $f(\boldsymbol{\upsilon}) = (f(\upsilon_1), \ldots, f(\upsilon_K))'$. Ainsi,

$$\mathbb{E}[C_t] = \vec{\mathfrak{u}}(t)\boldsymbol{\upsilon}, \tag{I.6.2}$$

où $\vec{\mathfrak{u}}(t)$ est remplacée par $\vec{\pi}$ lorsque la distribution stationnaire est choisie comme distribution initiale de la chaîne. Dans ce cas, l'espérance est constante en tout point de la chaîne

$$\mathbb{E}[C_t] = \vec{\pi}\boldsymbol{\upsilon}, \quad \forall\, t \in \mathbb{N}. \tag{I.6.3}$$

I.6.2. Fonction d'autocovariance

Puisqu'il est intéressant de comparer les fonctions d'autocovariance et d'autocorrélation des chaînes de Markov et des modèles de Markov à variables latentes, nous présentons ici les résultats concernant les chaînes de Markov et référons à la section 3.4 pour l'autocovariance des modèles de Markov à variables latentes.

Proposition I.6.1. *Sans supposer l'homogénéité, la fonction d'autocovariance d'une chaîne de Markov est donnée par*

$$\mathbb{C}ov[C_t, C_{t+h}] = \vec{\mathfrak{u}}(t)\boldsymbol{\Upsilon}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\upsilon} - (\vec{\mathfrak{u}}(t)\boldsymbol{\upsilon}) \times (\vec{\mathfrak{u}}(t+h)\boldsymbol{\upsilon}), \forall\, t,h \in \mathbb{N}, \tag{I.6.4}$$

$$= \vec{\pi}\boldsymbol{\Upsilon}\boldsymbol{\Gamma}^h\boldsymbol{\upsilon} - (\vec{\pi}\boldsymbol{\upsilon})^2, \quad \textit{si la chaîne est homogène et } \forall\, h \in \mathbb{N}. \tag{I.6.5}$$

En particulier, lorsque $h = 0$, nous obtenons la variance de C_t

$$\mathbb{V}ar[C_t] = \mathbb{C}ov[C_t, C_t] = \vec{\mathfrak{u}}(t)\boldsymbol{\Upsilon}\boldsymbol{\upsilon} - (\vec{\mathfrak{u}}(t)\boldsymbol{\upsilon})^2, \forall\, t,h \in \mathbb{N}, \tag{I.6.6}$$

$$= \vec{\pi}\boldsymbol{\Upsilon}\boldsymbol{\upsilon} - (\vec{\pi}\boldsymbol{\upsilon})^2, \quad \textit{si la chaîne est homogène et } \forall\, h \in \mathbb{N}. \tag{I.6.7}$$

Démonstration. Voir Appendice III.2. □

Proposition I.6.2. *Si la chaîne de Markov est ergodique et que sa matrice de transition est diagonalisable tel que défini à l'appendice I.3, alors la fonction d'autocorrélation d'une chaîne de Markov est donnée par*

$$\mathbb{C}\text{orr}[C_t, C_{t+h}] = \frac{\sum_{i=2}^{K} a_i b_i \lambda_i^h}{\sum_{i=2}^{K} a_i b_i}, \tag{I.6.8}$$

avec a_i, le i^e élément de $\vec{a} = \vec{\pi}\Upsilon Q$ et b_i, le i^e élément du vecteur $\mathbf{b} = Q^{-1}\boldsymbol{\upsilon}$. Cette seconde forme permet de bien illustrer le fait que la persistance de l'autocorrélation d'une chaîne de markov dépend de la seconde valeur propre de sa matrice de transition.

Démonstration. Voir appendice III.3. □

I.7. Variance de $g(X_t)$ par la formule de variance totale

La formule de variance totale est une alternative à la démarche présentée à la section 3.3 pour l'obtention de la $\mathbb{V}ar[g(X_t)]$.

Celle-ci stipule que si les réalisations $A_i \in \mathbf{A}$, $i = \{1,2,\ldots,n\}$ sont mutuellement exclusives et couvrent tout l'espace des réalisations possibles, alors la variance d'une variable aléatoire quelconque X est donnée par la formule suivante :

$$\begin{aligned}\mathbb{V}ar[X] = &\sum_{i=1}^{n} \mathbb{V}ar[X \mid A_i]\mathbb{P}[A_i] \\ &+ \sum_{i=1}^{n} \mathbb{E}[X \mid A_i]^2(1 - \mathbb{P}[A_i])\mathbb{P}[A_i] \\ &- 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathbb{E}[X \mid A_i]\mathbb{P}[A_i]\mathbb{E}[X \mid A_j]\mathbb{P}[A_j].\end{aligned} \tag{I.7.1}$$

Il est donc possible d'obtenir la variance de $g(X_t)$ grâce à cette formule de variance totale. En effet,

$$\begin{aligned}
\mathbb{V}ar[g(X_t)] &= \sum_{i=1}^{K} \mathbb{V}ar[g(X_t) \mid C_t = i]\mathbb{P}[C_t = i] \\
&\quad + \sum_{i=1}^{K} \mathbb{E}[g(X_t) \mid C_t = i]^2(1 - \mathbb{P}[C_t = i])\mathbb{P}[C_t = i] \\
&\quad - 2\sum_{i=2}^{K}\sum_{j=1}^{i-1} \mathbb{E}[g(X_t) \mid C_t = i]\mathbb{P}[C_t = i]\mathbb{E}[g(X_t) \mid C_t = j]\mathbb{P}[C_t = j], \\
&= \sum_{i=1}^{K} \sigma^2_{g|i}\ \mathfrak{u}_i(t) + \sum_{i=1}^{K} \mu^2_{g|i}(1 - \mathfrak{u}_i(t))\mathfrak{u}_i(t) - 2\sum_{i=2}^{K}\sum_{j=1}^{i-1} \mu_{g|i}\ \mathfrak{u}_i(t)\, \mu_{g|j}\ \mathfrak{u}_j(t), \qquad \text{(I.7.2)} \\
&= \vec{\mathfrak{u}}(t)\mathbf{\Sigma}^2_g\mathbf{1}_K + \vec{\mathfrak{u}}(t)\mathbf{M}^2_g\Big(\mathbf{1}_K - (\vec{\mathfrak{u}}(t))'\Big) - 2\Big\{\vec{\Omega}^{(t)}_g\Big[\Big(\mathbf{1}_K\vec{\Omega}^{(t)}_g\Big)\odot\Big(\mathbf{1}_K\mathbf{1}_K{}' - \llcorner_K\Big)\Big]\mathbf{1}_K\Big\}, \\
&= \vec{\mathfrak{u}}(t)\Big[\mathbf{\Sigma}^2_g + \mathbf{M}^2_g\Big]\mathbf{1}_K - \vec{\mathfrak{u}}(t)\mathbf{M}^2_g\ (\vec{\mathfrak{u}}(t))' - 2\Big\{\vec{\Omega}^{(t)}_g\Big[\Big(\mathbf{1}_K\vec{\Omega}^{(t)}_g\Big)\odot\Big(\mathbf{1}_K\mathbf{1}_K{}' - \llcorner_K\Big)\Big]\mathbf{1}_K\Big\}. \qquad \text{(I.7.3)}
\end{aligned}$$

avec $\odot$ correspondant au produit élément par élément (Produit d'Hadamard), $\mathbf{\Sigma}_g \overset{\text{def}}{=} \text{diag}(\boldsymbol{\sigma}_g)$, $\boldsymbol{\sigma}_g \overset{\text{def}}{=} \Big(\sigma_{g|1}, \ldots, \sigma_{g|K}\Big)$, $\vec{\Omega}^{(t)}_g \overset{\text{def}}{=} \vec{\mathfrak{u}}(t)\odot\boldsymbol{\mu}_g{}'$ et

$$\llcorner_K \overset{\text{def}}{=} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

une matrice triangulaire inférieure de dimension $K \times K$.

Démonstration. La preuve d'égalité entre la sommation et la forme matricielle du troisième terme de cette somme est en appendice III.4 □

Appendice II

Techniques et Résultats supplémentaires concernant les HMMs

II.1. Algorithme avant (*Forward algorithm*)

Comme le mentionnait la remarque 3.2.1, une solution possible au problème de filtrage découle directement de la proposition 3.2.1. Nous y avions déjà mentionné pouvoir adapter le résultat de la proposition afin de calculer la vraisemblance. En effet, à partir de ce résultat, nous pouvons faire la proposition suivante :

Proposition II.1.1. *La vraisemblance d'un modèle de Markov à variables latentes est obtenue en calculant la distribution marginale du vecteur $X_{1:T}$, soit*

$$\mathcal{L}_T(\boldsymbol{\theta}) = \vec{\mathfrak{u}}(1)\mathbf{P}(x_1)\left(\prod_{t=2}^{T}\boldsymbol{\Gamma}_{t-1}\mathbf{P}(x_t)\right)\mathbf{1}_K. \tag{II.1.1}$$

*De plus, puisque la multiplication matricielle est associative, il est possible d'évaluer ce produit grâce à un algorithme récursif connu sous le nom de l'**algorithme avant** et présenté à l'algorithme 3 : Algorithme avant* (p.140).

Remarque II.1.1. Il est important de souligner que l'algorithme avant permet d'obtenir la vraisemblance sous la spécification généralisée (p.28) dans laquelle la chaîne de Markov admet une dépendance entre C_t et $X_{1:t-1}$. C'est d'ailleurs sous cette généralisation que sera présentée la preuve de l'agorithme avant.

Démonstration. Tel que mentionné à la remarque II.1.1, nous faisons ici la preuve de l'algorithme avant sous la spécification généralisée (p.28).

Algorithme 3 : Algorithme avant

Soit les probabilités

$$\vec{\alpha}_t \overset{\text{def}}{=} (\mathbb{P}[X_{1:t}, C_t = 1], \ldots, \mathbb{P}[X_{1:t}, C_t = K]), \quad t = 1, \ldots, T, \tag{II.1.2}$$

habituellement appelées **probabilité vers l'avant** [a], avec

$$\vec{\alpha}_t(j) \overset{\text{def}}{=} \mathbb{P}[X_{1:t}, C_t = j], \tag{II.1.3}$$

$$= \left(\sum_{i=1}^{K} \vec{\alpha}_{t-1}(i) \gamma_{ij}(t-1) \right) p_j(x_t), \quad \forall\, j \in \mathcal{C}. \tag{II.1.4}$$

a. Lystig et Hughes [2002]

Sous forme matricielle, l'algorithme avant se résume à trois étapes :

Étape 1

$$\vec{\alpha}_1 \quad = \quad \vec{u}(1)\mathbf{P}(x_1)$$

Étape 2

Pour $t = 2, \ldots, T$, calculer les termes :

$$\vec{\alpha}_t = \vec{\alpha}_{t-1}\mathbf{\Gamma}_{t-1}\mathbf{P}(x_t)$$

Étape 3

$$\mathcal{L}_T(\boldsymbol{\theta}) \quad = \quad \vec{\alpha}_T \mathbf{1}_K.$$

Avec $\mathbf{P}(x_t) = \text{diag}(\vec{p}(x_t))$.

Coût computationnel	Référence
$\mathrm{O}(TK^2)$	Zucchini et Macdonald [2009](p. 60)

Remarque

Sujet au problème de soupassement arithmétique (*underflow*) et de dépassement d'entier (*overflow*).

Par la formule de la probabilité totale, nous avons dans un premier temps que la vraisemblance de ce modèle peut s'écrire sous la forme

$$\mathcal{L}_T(\boldsymbol{\theta}) \overset{\text{def}}{=} \mathbb{P}[x_{1:T}], \tag{II.1.5}$$

$$\overset{\text{PT}}{=} \sum_{j=1}^{K} \mathbb{P}[x_{1:T}, C_T = j]. \tag{II.1.6}$$

Le terme à l'intérieur de la somme peut ensuite être décomposé de la manière suivante :

$$\begin{aligned}\mathbb{P}[x_{1:T}, C_T = j] &\overset{\text{PT}}{=} \sum_{i=1}^{K} \mathbb{P}[x_{1:T}, C_T = j, C_{T-1} = i], \\ &\overset{\text{Bayes}}{=\!=} \sum_{i=1}^{K} \mathbb{P}[x_T \mid C_T = j, \cancel{C_{T-1} = i}, x_{1:T-1}]\mathbb{P}[x_{1:T-1}, C_T = j, C_{T-1} = i], \\ &\overset{\text{Bayes}}{=\!=} \sum_{i=1}^{K} \mathbb{P}[x_T \mid C_T = j, x_{1:T-1}]\mathbb{P}[C_T = j \mid C_{T-1} = i, x_{1:T-1}]\, \mathbb{P}[x_{1:T-1}, C_{T-1} = i]\,.\end{aligned}$$

Ainsi, nous reconnaissons dès lors la relation entre les termes en surbrillance. En notant

$$\vec{\alpha}_t(j) \overset{\text{def}}{=} \mathbb{P}[X_{1:t}, C_t = j], \tag{II.1.7}$$

nous avons que

$$\vec{\alpha}_T(j) = \sum_{i=1}^{K} p_j(x_T \mid x_{1:T-1}) \Upsilon_{ij}(T-1)\, \vec{\alpha}_{T-1}(i) \tag{II.1.8}$$

et donc que

$$\vec{\alpha}_t(j) = \sum_{i=1}^{K} \vec{\alpha}_{t-1}(i)\, \Upsilon_{ij}(t-1) p_j(x_t \mid x_{1:t-1}), \quad \begin{cases} \forall\, j \in \{1, \ldots, K\} \text{ et} \\ \forall\, t \in \{2, \ldots, T\}. \end{cases} . \tag{II.1.9}$$

Cette relation tient jusqu'en $t = 1$, avec ces derniers termes obtenus directement par définition et le théorème de Bayes :

$$\vec{\alpha}_1(j) \overset{\text{def}}{=} \mathbb{P}[x_1, C_1 = j] \overset{\text{Bayes}}{=\!=} \mathbb{P}[C_1 = j]\mathbb{P}[x_1 \mid C_1 = j\,] = \mathfrak{u}_j(1) p_j(x_1), \quad \forall\, j \in \{1, \ldots, K\}. \tag{II.1.10}$$

Finalement, sous forme matricielle, ces relations sont équivalentes à

$$\vec{\alpha}_t = \vec{\alpha}_{t-1} \mathbf{\Gamma}_{t-1}\, \vec{p}(x_t \mid x_{1:t-1}\,) \tag{II.1.11}$$

et

$$\vec{\alpha}_1 = \vec{\mathfrak{u}}(1) \vec{p}(x_1). \tag{II.1.12}$$

□

Bien que cette technique soit en théorie très efficace pour obtenir une expression de la vraisemblance, elle souffre malheureusement du problème de soupassement arithmétique et devra être adaptée. C'est pourquoi nous avons choisi de présenter plutôt le filtre d'Hamilton, une solution élégante à ce problème et un algorithme adaptatif très important dans l'étude des modèles de Markov à variables latentes.

II.2. Inférence conditionnelle à $\boldsymbol{\mathcal{F}}_T$

Telle que mentionné, une autre inférence d'intérêt consiste à prévoir les états cachés de la chaîne cachée en conditionnant cette fois sur l'entièreté de la série d'observations. Deux approches sont étudiées ici. D'abord, nous présenterons le **filtre de Kim** (Kim [1994]) permettant une inférence locale en calculant $\mathbb{P}[C_t = i \mid x_{1:T}]$, $\forall\, i \in \mathcal{C}$ et $\forall\, t \in \mathbb{N}^\star$. Nous aborderons ensuite l'**algorithme de Viterbi**, permettant une inférence globale en obtenant la séquence d'états inférés $\hat{C}_{1:T} \stackrel{\text{def}}{=} \arg\max_{c_{1:T}} \mathbb{P}[C_{1:T} = c_{1:T} \mid x_{1:T}]$.

II.2.1. Filtre de Kim

Aussi dénommé l'**algorithme de lissage arrière** (*backward smoothing algorithm*), le filtre de Kim permet de complémenter l'information obtenue grâce au filtre d'Hamilton. Alors que ce dernier permet une inférence dite vers l'avant en calculant les termes $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_t]$ pour $t = 1, \ldots, T$, le filtre de Kim permet une inférence dite vers l'arrière en obtenant les termes $\mathbb{P}[C_t \mid \boldsymbol{\mathcal{F}}_T]$ pour $t = T-1, \ldots, 1$. Ces probabilités—appelées **probabilités lissées** (*smoothed probabilities*)—sont calculées récursivement à partir des probabilités obtenues par la méthode du filtre d'Hamilton.

Il est important de rappeler que l'on conditionne ici sur l'état suivant de la chaîne—soit en $t+1$—lorsque l'on fait l'inférence sur l'état en t. En reprenant les termes calculés lors du filtre d'Hamilton, le filtre de Kim n'est autre chose qu'une application de la formule des probabilités totales et du théorème de Bayes.

Le premier terme est obtenu directement du filtre d'Hamilton, soit

$$\vec{\omega}_{T\,|\,T}(i) \stackrel{\text{def}}{=} \mathbb{P}[C_T = i \mid x_{1:T}] = \vec{\omega}_T(i). \tag{II.2.1}$$

Pour les termes suivants, nous avons que

$$\vec{\omega}_{t\,|\,T}(i) \stackrel{\text{def}}{=} \mathbb{P}[C_t = i \mid x_{1:T}] = \sum_{j=1}^{K} \mathbb{P}[C_t = i, C_{t+1} = j \mid x_{1:T}], \tag{II.2.2}$$

avec

$$\begin{aligned}
\mathbb{P}[C_t = i, C_{t+1} = j \mid x_{1:T}] &= \mathbb{P}[C_{t+1} = j \mid x_{1:T}]\mathbb{P}\Big[C_t = i \;\Big|\; C_{t+1} = j, \boxed{x_{1:T}}\Big], \\
&= \vec{\omega}_{t+1\,|\,T}(j)\mathbb{P}\Big[C_t = i \;\Big|\; C_{t+1} = j, \boxed{x_{1:t}}\Big], \\
&= \vec{\omega}_{t+1\,|\,T}(j)\frac{\mathbb{P}[C_t = i, C_{t+1} = j \mid x_{1:t}]}{\mathbb{P}[C_{t+1} = j \mid x_{1:t}]}, \\
&= \vec{\omega}_{t+1\,|\,T}(j)\frac{\mathbb{P}[C_t = i \mid x_{1:t}]\mathbb{P}[C_{t+1} = j \mid C_t = i, x_{1:t}]}{\sum_{i=1}^{K}\mathbb{P}[C_t = i, C_{t+1} = j \mid x_{1:t}]}, \\
&= \vec{\omega}_{t+1\,|\,T}(j)\frac{\vec{\omega}_t(i)\mathbb{P}[C_{t+1} = j \mid C_t = i, x_{1:t}]}{\sum_{k=1}^{K}\vec{\omega}_t(k)\mathbb{P}[C_{t+1} = j \mid C_t = k, x_{1:t}]}
\end{aligned} \tag{II.2.3}$$

et où le terme $\vec{\omega}_{t+1\,|\,T}(j)$ provient de l'étape précédente de l'algorithme. La preuve de la transition en surbrillance est en appendice III.5.

Une version vectorielle de cet algorithme, inspirée des travaux de Hamilton [1994](p. 694), est présentée à l'algorithme 4 : Filtre de Kim (p.144).

II.2.2. Algorithme avant-arrière

Puisque le filtre de Kim nécessite d'abord l'utilisation du filtre d'Hamilton, cette combinaison de méthodes porte un nom, l'**algorithme avant-arrière** (*forward-backward algorithm* ou *forward filtering-backward smoothing algorithm*). C'est dans les travaux de Baum [1972] qu'est d'abord proposée cette combinaison.

La réalisation de cet algorithme, à l'instar de ses composantes, nécessite que nous connaissions trois distributions :

$$\begin{aligned}
\text{Distribution initiale} \quad & \vec{\mathbf{u}}(1), \\
\text{Probabilités de transition} \quad & \mathbf{\Gamma}_t = (\gamma_{ij}(t))_{i,j\in\mathcal{C}}, \\
\text{Probabilités d'émission} \quad & \mathbb{P}[X_t \mid C_t].
\end{aligned}$$

Afin de bien cerner l'intérêt de cet algorithme—un exemple simple de programmation dynamique—il suffit de nous intéresser à ce que chacune des parties contribue. Sans nous éterniser dans les détails, les filtres d'Hamilton et de Kim, une fois combinés, nous permettent d'obtenir les informations répertoriées dans la figure II.1.

II.2.3. Algorithme de Viterbi

Alors que l'algorithme avant-arrière permet une inférence dite locale, certains algorithmes permettent de faire une inférence dite globale. Dans notre situation, ceci correspond à obtenir

Algorithme 4 : Filtre de Kim

Soit les notations suivante, et leurs interprétation probabilistes :

$$\vec{\omega}_{t\,|\,t} = \vec{\omega}_t = (\mathbb{P}[C_t = 1 \mid \boldsymbol{\mathcal{F}}_t], \dots, \mathbb{P}[C_t = K \mid \boldsymbol{\mathcal{F}}_t]), \tag{II.2.4}$$

$$\vec{\omega}_{t+h\,|\,t} = (\mathbb{P}[C_{t+h} = 1 \mid \boldsymbol{\mathcal{F}}_t], \dots, \mathbb{P}[C_{t+h} = K \mid \boldsymbol{\mathcal{F}}_t]), \tag{II.2.5}$$

$$\vec{\omega}_{t\,|\,T} = (\mathbb{P}[C_t = 1 \mid \boldsymbol{\mathcal{F}}_T], \dots, \mathbb{P}[C_t = K \mid \boldsymbol{\mathcal{F}}_T]). \tag{II.2.6}$$

Étape 1

Directement du filtre d'Hamilton, nous obtenons le premier terme :

$$\vec{\omega}_{T\,|\,T} = \mathbb{P}[C_T \mid \boldsymbol{\mathcal{F}}_T] = \vec{\omega}_T \tag{II.2.7}$$

Étape 2

Pour $t = T-1, \dots, 1$, calculer les termes :

$$\vec{\omega}_{t\,|\,T} = \vec{\omega}_{t\,|\,t} \odot \left\{ \left[\vec{\omega}_{t+1\,|\,T} \oslash \vec{\omega}_{t+1\,|\,t} \right] \boldsymbol{\Gamma}_t' \right\}, \tag{II.2.8}$$

$$= \vec{\omega}_{t\,|\,t} \odot \left\{ \left[\vec{\omega}_{t+1\,|\,T} \oslash \left(\vec{\omega}_{t\,|\,t} \boldsymbol{\Gamma}_t \right) \right] \boldsymbol{\Gamma}_t' \right\} \tag{II.2.9}$$

où $\odot$ et $\oslash$ correspondent respectivement au produit et à la division de matrice terme par terme de Hadamard.

Coût computationnel

$\mathrm{O}(TK^2)$

Référence

Kim [1994](p. 9) et
Hamilton [1994](p. 694)

Remarque

Cet algorithme est valide même pour un modèle de Markov à variables latentes non-homogène tel que défini à la spécification généralisée (p.28).

Filtre d'Hamilton	Densité conditionnelle	$\mathbb{P}[X_t \mid x_{1:t-1}]$	a_t	$\forall\, t \in 1, \dots, T$
	Probabilité filtrée	$\mathbb{P}[C_t \mid x_{1:t}]$	$\vec{\omega}_t$	
	Probabilité prédictive	$\mathbb{P}[C_{t+1} \mid x_{1:t}]$	$\vec{\omega}_{t+1\,|\,t}$	
Filtre de Kim	Probabilité lissée	$\mathbb{P}[C_t \mid x_{1:T}]$	$\vec{\omega}_{t\,|\,T}$	$\forall\, t \in T-1, \dots, 1$

Figure II.1 – Probabilités obtenues grâce à l'algorithme avant-arrière

la suite d'état $c^\star_{1:T}$ telle que

$$c^\star_{1:T} = \arg\max_{c_{1:T}} \mathbb{P}[c_{1:T} \mid x_{1:T}], \tag{II.2.10}$$

ou de manière équivalente avec la probabilité conjointe

$$c^{\star}_{1:T} = \arg\max_{c_{1:T}} \mathbb{P}[x_{1:T}, c_{1:T}]. \tag{II.2.11}$$

Théoriquement, la solution à ce problème pourrait être obtenue en évaluant $\mathbb{P}[x_{1:T}, c_{1:T}]$ pour chacune des K^T trajectoires possibles pour $c_{1:T}$ et en choisissant celle qui maximise cette probabilité. Cependant, tel que démontré à la section 3.5.1, ceci est simplement impraticable, même pour une chaîne relativement courte sur un espace d'états de cardinalité $K = 2$.

Un algorithme connu sous le nom de l'**algorithme de Viterbi** permet de contourner ce problème et d'y trouver une solution en $\mathrm{O}(TK^2)$ opérations. Cet algorithme, un analogue de l'**algorithme de propagation des convictions** (*Belief propagation*), est d'abord présenté par Viterbi [1967] dans le but de borner la probabilité d'erreur lors du décodage de messages, puis adapter par Forney [1973] à l'estimation de la séquence d'état d'une chaîne de Markov à temps discret sur un espace d'états fini.

À la manière de Forney [1973], et afin de simplifier la présentation des étapes de l'algorithme, nous proposons la notation ci-bas, ainsi que la méthode pour obtenir ces termes récursivement.

Proposition II.2.1. *Soit les probabilités suivantes et la notation associée :*

$$\xi_1(i) \stackrel{\text{def}}{=} \mathbb{P}[x_1, C_1 = i], \quad i = 1, \ldots, K, \tag{II.2.12}$$

$$\xi_t(i) \stackrel{\text{def}}{=} \max_{c_{1:t-1}} \mathbb{P}[x_{1:t}, c_{1:t-1}, C_t = i], \quad i = 1, \ldots, K \text{ et } t = 2, \ldots, T. \tag{II.2.13}$$

Il est possible d'obtenir le terme $\xi_{t+1}(j)$ grâce à la relation recursive suivante :

$$\xi_{t+1}(j) = \mathbb{P}[\, x_{t+1} \mid x_{1:t}, C_{t+1} = j\,] \times \max_{i=1,\ldots,K} \{\mathbb{P}[C_{t+1} = j \mid C_t = i, x_{1:t}]\xi_t(i)\}, \tag{II.2.14}$$

$$= p_j(\, x_{t+1} \mid x_{1:t}\,) \times \max_{i=1,\ldots,K} \{\Upsilon_{ij}(t) \cdot \xi_t(i)\}. \tag{II.2.15}$$

Remarque II.2.1 (Soupassement arithmétique). Tout comme les probabilités calculées dans le cadre de l'algorithme avant, les termes de la proposition II.2.1 sont sensibles au soupassement arithmétique.

Afin de rémédier à ce problème, il est possible de normaliser les termes un pas de temps à la fois. Ainsi, nous aurons que

$$\xi_t(i) \leftarrow \frac{\xi_t(i)}{\sum_{i=1}^{K} \xi_t(i)}. \tag{II.2.16}$$

Il semble déjà à cette étape que la maximisation pourra être effectuée un pas à la fois. L'algorithme de Viterbi est présenté à l'algorithme 5 : Algorithme de Viterbi (p.146) et les démonstrations de celui-ci et de la proposition II.2.1 sont reportées en appendice III.6.

Algorithme 5 : Algorithme de Viterbi

Soit la suite d'états $i_{1:T}$ satisfaisant la relation

$$i_{1:T} = \arg\max_{c_{1:T}} \mathbb{P}[c_{1:T} \mid x_{1:T}]. \tag{II.2.17}$$

Cette séquence est obtenue grâce à l'algorithme récursif suivant :

Étape 1

Calculer les termes $\xi_t(i)$, $i = 1, \ldots, K$, en suivant la méthode récursive présentée à la proposition II.2.1 pour $t = 1, \ldots, T$.

Étape 2

Calculer :

$$i_T \stackrel{\text{def}}{=} \arg\max_{i=1,\ldots,K} \xi_T(i). \tag{II.2.18}$$

Étape 3

Pour $t = T-1, \ldots, 1$, calculer les termes :

$$\begin{aligned} i_t &\stackrel{\text{def}}{=} \arg\max_{j=1,\ldots,K} \{\mathbb{P}[C_{t+1} = i_{t+1} \mid C_t = j, x_{1:t}] \cdot \xi_t(j)\}, \\ &= \arg\max_{j=1,\ldots,K} \left\{ \gimel_{j(i_{t+1})}(t) \cdot \xi_t(j) \right\}. \end{aligned} \tag{II.2.19}$$

Coût computationnel	Référence
$\mathrm{O}(TK^2)$	Augustyniak [2018](p. 97)

Remarque

Les termes provenant de la proposition II.2.1 et calculés à l'étape 1 peuvent être emmagasinés dans une matrice de dimensions $T \times K$, facilitant la manipulation de ceux-ci. C'est cette caractéristique qui fait de l'algorithme de Viterbi un autre exemple de programmation dynamique.

II.3. Estimation des paramètres

Estimer les paramètres d'un modèle de Markov à variables latentes peut s'avérer très complexe. Plusieurs considérations techniques doivent être prises en compte afin de nous assurer entre autres que le modèle est identifiable et que l'EMV existe, de construire correctement la vraisemblance, de choisir la bonne méthode d'estimation, d'assurer que le vecteur de paramètres obtenu maximise la vraisemblance (i.e. que c'est un EMV) et que l'EMV

trouvé est fortement convergent et asymptotiquement efficace (Ces derniers termes seront définis dans les sections II.3.2 et II.3.6 respectivement).

Par ailleurs, les paramètres tels que définis pour le moment sont contraints. Il est alors habituellement suggéré de reparamétriser le modèle afin d'accélérer ou même de rendre possible le processus d'optimisation. Cette manoeuvre ayant des répercussions sur la majorité des quantités calculées dans le contexte du processus d'optimisation, nous réservons la section II.3.5 afin de l'étudier et de présenter à nouveau les principaux résultats devant être ajustés.

II.3.1. Idenfiabilité du modèle

Définition II.3.1. *Une famille de distribution* $\{\mathbb{P}[x \mid \theta] : \theta \in \boldsymbol{\Theta}\}$ *est dites identifiable si* $\forall\, \theta, \theta' \in \boldsymbol{\Theta}$*, nous avons que*

$$\mathbb{P}[x \mid \theta] = \mathbb{P}[x \mid \theta'] \text{ pour (presque) tout } x \in \chi \implies \theta = \theta', \tag{II.3.1}$$

avec $\boldsymbol{\Theta}$ *correspondant à l'espace des paramètres. Si cette condition n'est pas respectée, alors* $\exists \tilde{\theta} \in \boldsymbol{\Theta}$ *tel que l'ensemble*

$$\mathcal{U}\big(\tilde{\theta}\big) \stackrel{\text{def}}{=} \Big\{\theta' \in \boldsymbol{\Theta} \colon \mathbb{P}[x \mid \theta'] = \mathbb{P}\Big[x \;\Big|\; \tilde{\theta}\Big] \text{ pour (presque) tout } x \in \chi\Big\}, \tag{II.3.2}$$

contient plus d'un élément. Ceci a pour effet que la famille n'est plus identifiable, et qu'il n'est pas garanti que les paramètres du modèle puisse être identifiés adéquatement par rapport à un échantillon donné. Plusieurs combinaisons distinctes de paramètres seraient alors associées à une même valeur de la vraisemblance.

Sans nous éterniser sur le sujet, il est évidemment important que le modèle retenu soit identifiable, ou du moins qu'il soit possible d'ajouter des contraintes sur les paramètres et leurs intéractions afin de le rendre identifiable. Tel qu'il a été défini jusqu'ici, notre modèle n'est pas strictement identifiable selon la définition II.3.1. Par exemple, il est clair à première vue que l'ordre des distributions du modèles n'a pas été une préoccupation jusqu'à ce point. Nous nous sommes contentés pour le moment d'étiquetter les régimes en utilisant les entiers positifs, sans que ces nombres aient de signification intrinsèque dans notre modèle. Dans sa forme actuelle, notre modèle admet donc un ensemble $\mathcal{U}\big(\tilde{\theta}\big)$ de cardinalité supérieur ou égale à $K!$ puisque chacune des K distributions peut se retrouver à n'importe quelle position relative aux autres sans affecter la vraisemblance.

Afin de rendre un modèle de Markov à variables latentes général identifiable, les trois contraintes suivantes suffisent habituellement :

(1) La chaîne de Markov doit être irréductible[1], empêchant ainsi qu'un régime ne soit inatteignable et que les paramètres de la distribution correspondante puissent alors prendre n'importe quelle valeur dans $\boldsymbol{\Theta}$, sans affecter la vraisemblance.
(2) Aucunes deux composantes du modèle ne peuvent partager tous les paramètres de leurs lois respectives. Ceci revient à dire que $\forall\, i \neq j \in \mathcal{C}$, θ_i doit différer que θ_j pour au moins un des paramètres de la distribution.
(3) Les vecteurs de paramètres $\{\theta_i\}_{i=1,\dots,K}$ doivent se voir imposer une relation d'ordre quelconque afin que leur position soit fixe l'un par rapport à l'autre. Lorsqu'il est question de paramètres univariées, il est possible d'imposer directement $\theta_1 < \theta_2 < \cdots < \theta_K$. Sinon, il faut simplement choisir un des paramètres sur lequel imposer cette relation d'ordre.

Remarque II.3.1. Même si en pratique il se peut qu'aucun paramètre ne soit distinct pour toutes les composantes du modèle et que nous ne puissions donc pas choisir un seul d'entre-eux pour ordonner strictement les composantes, ce problème est facilement contourné en utilisant un deuxième paramètre afin d'ordonner les composantes pour lesquelles le paramètre choisi prend la même valeur.

Finalement, certaines familles de distributions ne sont pas identifiables, même sous les conditions énumérées plus haut. Ce n'est heureusement pas le cas des familles gaussiennes ou de poisson qui sont habituellement réputées identifiables. Pour plus de détails concernant les autres familles de distribution ainsi que quelques exemples illustrant les limites des trois contraintes présentées ci-haut, nous référons les intéressé.e.s à la section 3.3.2 de Augustyniak [2018](p. 40-43).

Remarque II.3.2 (*label-switching problem*). Si une approche baysienne était retenue pour l'estimation des paramètres, l'interchangeabilité des paramètres deviendrait problèmatique et la contraintes (3) ne suffirait plus. Ce problème, connu sous le nom du *label-switching problem*, est d'abord mentionné dans les travaux de Redner et Walker [1984] pour identifier le problème d'invariance de la vraisemblance sous une permutation des étiquettes. Nous n'explorerons pas cet enjeux en plus de détails ici, mais plus d'information ainsi que quelques solutions possibles à ce problème peuvent entre autres être retrouvées dans les travaux de Diebolt et Robert [1994], Celeux [1998], Stephens [2000], Jasra, Holmes et Stephens [2005] et Puolamäki et Kaski [2009].

Il est à noter que les modèles de mélange présentés à la section 1.3 souffrent eux aussi pour les mêmes raisons de ce problème d'identifiabilité. Les contraintes proposées pour les

1. Voir définition I.2.1

HMMs sont généralement suffisantes pour les modèles de mélange, à l'exception de la première contrainte qui est remplacée par la contrainte que les probabilités de mélange soient strictement positives, $\vec{\pi} > 0$.

II.3.2. Existence de l'EMV

Plusieurs facteurs, dont une vraisemblance non bornée, peuvent faire en sorte que l'EMV n'existe pas au sens stricte. Ceci peut se produire lorsque, par exemple, la volatilité d'une distribution normale tend vers 0. Puisque ceci correspond à la frontière de $\boldsymbol{\Theta}$ et à une distribution dites dégénérée [2] exclue la majorité du temps, une solution possible est de restreindre le domaine des paramètres et imposer la condition

$$\min_{j,k\in\{1,\dots,K\}} \frac{\sigma_k}{\sigma_j} \geq c > 0 \tag{II.3.3}$$

proposée par Hathaway [1985], où les σ_i, $i \in \{1, \dots, K\}$ correspondent aux écarts-types des K composantes normales du mélange. Il est aussi démontré dans ce dernier ouvrage que la condition II.3.3 implique que l'EMV est alors fortement convergent (*strongly consistent*), c'est-à-dire que $\widehat{\boldsymbol{\theta}} \overset{p.s.}{\to} \boldsymbol{\theta}$. Nous présenterons une autre condition permettant d'affirmer que l'EMV est fortement convergent au théorème II.3.2 de la section II.3.6.

II.3.3. Obtenir une expression analytique pour la vraisemblance

Une des principales considérations concernant la vraisemblance est à savoir si celle-ci peut être obtenue analytiquement. Calculer la vraisemblance pour un modèle de Markov à variables latentes sous les hypothèses du modèle général de la spécification généralisée (p.28) est l'un des cas les plus simple. En effet, le filtre d'Hamilton présenté à la section 3.6.1.1 nous donne directement une forme analytique pour la vraisemblance du modèle, permettant une maximisation numérique par descente de gradient, sans avoir recours à des simulations.

Nous utiliserons l'outil d'optimisation sans contrainte `optim` du language `R` afin d'optimiser notre modèle. Pour ce faire, nous aurons besoin de pouvoir calculer le gradient de la vraisemblance, soit le score, ainsi que la matrice hessienne et son inverse additif, la matrice d'information observée.

Remarque II.3.3. Si toutefois la vraisemblance ne peut être obtenue analytiquement, l'utilisation de l'algorithm EM—ou de sa version spécifique aux *HMMs*, l'algorithme de Baum-Welch—est nécessaire afin d'obtenir l'EMV grâce à une procédure itérative.

2. La loi de probabilité d'une variable aléatoire $x \in \chi$ est dites dégénérée si $\exists a \in \chi$ tel que $\mathbb{P}[a] = 1$.

II.3.4. Gradient, matrice hessienne et d'information observée

Définition II.3.2 (Gradient de la log-vraisemblance)**.** *Soit* $\ell : \boldsymbol{\Theta} \mapsto \mathbb{R}$*, la fonction de log-vraisemblance* $\ell_T(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T}))$*, où* $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\theta_1, \ldots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$*, avec* D *le nombre de paramètres libres du modèle et* $\boldsymbol{\Theta}$ *l'espace des paramètres. Le* ***gradient*** *de* ℓ *en* $\boldsymbol{\theta}$*, noté* $\boldsymbol{\nabla}\ell_T(\boldsymbol{\theta})$ *— aussi appelé le* ***score*** *et noté* $\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})$ *— est le vecteur des dérivées partielles de* $\ell_T(\boldsymbol{\theta})$*,*

$$\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T}) \stackrel{\text{def}}{=} \boldsymbol{\nabla}\ell_T(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial \ell_T}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \ell_T}{\partial \theta_1}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial \ell_T}{\partial \theta_D}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \theta_1}\log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})) \\ \vdots \\ \frac{\partial}{\partial \theta_D}\log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})) \end{pmatrix} \in \mathbb{R}^D. \tag{II.3.4}$$

Ce vecteur colonne est fonction de $\boldsymbol{\theta}$ *et de l'échantillon* $x_{1:T}$*.*

Le gradient de la log-vraisemblance à plus d'une utilité lors de l'estimation des paramètres de notre modèle. Dans un premier temps, donner explicitement la forme de ce gradient à la routine d'optimisation permet potentiellement de diminuer le temps nécessaire à sa convergence vers l'EMV. De plus, c'est ce gradient qui nous permet de nous assurer que l'algorithme à bel et bien convergé vers le maximum de la fonction et que l'estimateur trouvé est celui qui est fortement convergent et asymptotiquement efficace [3].

Définition II.3.3 (Matrice Hessienne de la log-vraisemblance)**.** *Soit* $\ell : \boldsymbol{\Theta} \mapsto \mathbb{R}$*, la fonction de log-vraisemblance* $\ell_T(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T}))$*, où* $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\theta_1, \ldots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$*, avec* D *le nombre de paramètres libres du modèle. La* ***matrice hessienne*** *de* ℓ *en* $\boldsymbol{\theta}$*, noté* $\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta})$*, est la matrice carrée symétrique des dérivées partielles secondes de* $\ell_T(\boldsymbol{\theta})$*,*

$$\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial^2 \ell_T}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\boldsymbol{\theta}) = \frac{\partial^2 \ell_T}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}(\boldsymbol{\theta}) \in \mathbb{R}^{D \times D}. \tag{II.3.5}$$

L'élément au croisement de la i^e *ligne et de la* j^e *colonne de la matrice* $\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta})$ *est noté* $[\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta})]_{ij}$ *et correspond à*

$$[\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta})]_{ij} = \frac{\partial^2 \ell_T}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) = \frac{\partial^2 \ell_T}{\partial \theta_j \partial \theta_i}(\boldsymbol{\theta}). \tag{II.3.6}$$

La matrice $\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta})$ *est fonction de* $\boldsymbol{\theta}$ *et de* $x_{1:T}$*.*

3. Voir section II.3.6

Définition II.3.4 (Matrice d'information observée)**.** *La* ***matrice d'information observée****, notée* $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid x_{1:T})$*, est définie comme le négatif de la matrice hessienne de la log-vraisemblance par rapport au vecteur* $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\theta_1, \ldots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$*, soit*

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid x_{1:T}) \stackrel{\text{def}}{=} -\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell_T}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\boldsymbol{\theta}), \tag{II.3.7}$$

avec l'élément au croisement de la i^{e} *ligne et de la* j^{e} *colonne, noté* $\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid x_{1:T})\right]_{ij}$*, correspondant à*

$$\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid x_{1:T})\right]_{ij} = -\left[\boldsymbol{\mathcal{H}}_{\ell_T}(\boldsymbol{\theta})\right]_{ij} = -\frac{\partial^2 \ell_T}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell_T}{\partial \theta_j \partial \theta_i}(\boldsymbol{\theta}). \tag{II.3.8}$$

La matrice $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid x_{1:T})$ *est fonction de* $\boldsymbol{\theta}$ *et de* $x_{1:T}$*.*

Remarque II.3.4. Quoique la matrice d'information observée corresponde toujours au négatif de la matrice hessienne de la log-vraisemblance, la matrice hessienne—comme le gradient—peut être définie par rapport à n'importe quelle fonction dérivable.

Obtenir ces trois expressions n'est certe pas trivial, mais Lystig et Hughes [2002] proposent une technique récursive très élégante permettant d'obtenir celles-ci simultanément et à faible coût en terme de complexité. Ce qui rend cet algorithme intéressant est qu'il peut être intégré directement à la récursion du filtre d'Hamilton, permettant de construire le gradient et la matrice hessienne au fur et à mesure que l'on construit la vraisemblance. Cependant, ce qui permet à cette technique de se démarquer est le fait qu'elle permet d'obtenir la valeur exacte de la matrice d'information observée—en plus du score et de la log-vraisemblance—, alors que, d'après les auteurs, ses concurrentes ne permettent habituellement qu'une estimation de celle-ci.

II.3.4.1. Algorithme récursif pour le calcul du score

Proposition II.3.1 (Algorithme récursif pour le score)**.** *Soit le vecteur de paramètres libres du modèle* $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\theta_1, \ldots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$ *et la notation pour les dérivées partielles de* $\mathbb{P}[X_{1:t}, C_t = j]$ *mises à l'échelle*

$$\psi_t(j; \theta_d) \stackrel{\text{def}}{=} \frac{\frac{\partial}{\partial \theta_d} \mathbb{P}[X_{1:t}, C_t = j]}{\mathbb{P}[X_{1:t-1}]} = \frac{\frac{\partial}{\partial \theta_d} \vec{\alpha}_t(j)}{\mathbb{P}[X_{1:t-1}]} \quad \begin{cases} \forall\, j \in \{1, \ldots, K\}, \\ \forall\, d \in \{1, \ldots, D\} \text{ et} \\ \forall\, t \in \{1, \ldots, T\}, \end{cases} \tag{II.3.9}$$

avec les probabilités vers l'avant

$$\vec{\alpha}_t(j) \stackrel{\text{def}}{=} \mathbb{P}[X_{1:t}, C_t = j], \tag{II.3.10}$$

$$= \sum_{i=1}^{K} \vec{\alpha}_{t-1}(i) \cdot \Upsilon_{ij}(t-1) \cdot p_j(x_t), \quad \forall\, j \in \mathcal{C} \tag{II.3.11}$$

provenant de l'algorithme avant.

Soit les notations suivantes, et leurs interprétation probabilistes provenant du filtre d'Hamilton :

$$\vec{a}_t \stackrel{\text{def}}{=} (\mathbb{P}[X_t, C_t = 1 \mid X_{1:t-1}], \ldots, \mathbb{P}[X_t, C_t = K \mid X_{1:t-1}]), \tag{II.3.12}$$

$$a_t \stackrel{\text{def}}{=} \mathbb{P}[X_t \mid X_{1:t-1}], \tag{II.3.13}$$

$$\vec{\omega}_t \stackrel{\text{def}}{=} (\mathbb{P}[C_t = 1 \mid X_{1:t}], \ldots, \mathbb{P}[C_t = K \mid X_{1:t}]). \tag{II.3.14}$$

Les termes de la récursion peuvent alors être obtenus récursivement par les relations

$$\begin{aligned}\psi_t(j;\theta_d) = \Bigg[\sum_{i=1}^{K} \Bigg\{ & \psi_{t-1}(i;\theta_d) \cdot \Upsilon_{ij}(t-1) \cdot p_j(x_t) \\ & + \vec{a}_{t-1}(i) \left[\frac{\partial \Upsilon_{ij}}{\partial \theta_d}(t-1) \right] p_j(x_t) \\ & + \vec{a}_{t-1}(i) \cdot \Upsilon_{ij}(t-1) \left[\frac{\partial p_j}{\partial \theta_d}(x_t) \right] \Bigg\} \Bigg] \times (a_{t-1})^{-1}, \quad \begin{cases} \forall\, j \in \{1, \ldots, K\}, \\ \forall\, d \in \{1, \ldots, D\} \ et \\ \forall\, t \in \{2, \ldots, T\}, \end{cases}\end{aligned} \tag{II.3.15}$$

et initialisées en $t = 1$ par les relations

$$\psi_1(j;\theta_d) \stackrel{\text{def}}{=} \frac{\partial}{\partial \theta_d} \mathbb{P}[X_1, C_1 = j], \tag{II.3.16}$$

$$= \left[\frac{\partial \mathfrak{u}_j}{\partial \theta_d}(1) \right] p_j(x_1) + \mathfrak{u}_j(1) \left[\frac{\partial p_j}{\partial \theta_d}(x_1) \right], \quad \begin{cases} \forall\, j \in \{1, \ldots, K\} \ et \\ \forall\, d \in \{1, \ldots, D\}, \end{cases} \tag{II.3.17}$$

avec

$$\left[\frac{\partial \Upsilon_{iK}}{\partial \theta_d}(t-1) \right] = \left(- \sum_{j=1}^{K-1} \left[\frac{\partial \Upsilon_{ij}}{\partial \theta_d}(t-1) \right] \right), \quad \forall\, i \in \{1, \ldots, K\} \tag{II.3.18}$$

et

$$\left[\frac{\partial \mathfrak{u}_K}{\partial \theta_d}(1) \right] = \left(- \sum_{i=1}^{K-1} \left[\frac{\partial \mathfrak{u}_i}{\partial \theta_d}(1) \right] \right). \tag{II.3.19}$$

Le score en θ_d, $\forall\, d \in \{1,\ldots,D\}$, est finalement donné par

$$\boldsymbol{\mathcal{S}}(\theta_d \mid x_{1:T}) \overset{\text{def}}{=} \frac{\partial \ell_T}{\partial \theta_d}(\boldsymbol{\theta}) = \frac{\Psi_T(\theta_d)}{a_T}, \tag{II.3.20}$$

avec $\Psi_T(\theta_d) = \sum_{i=1}^{K} \psi_T(i;\theta_d)$.

Alternativement, il est possible d'exprimer ces relations sous forme vectorielle, facilitant l'exécution de cet algorithme d'un point de vue computationnel et rapprochant celui-ci de l'algorithme du filtre d'Hamilton présenté à l'algorithme algorithme 1 : Filtre d'Hamilton (p.38)*. Une version vectorielle de cet algorithme est présentée à l'algorithme algorithme 6 : Calcul récursif du score* (p.154)*.*

Démonstration. Soit le vecteur de paramètres libres du modèle $\boldsymbol{\theta} \overset{\text{def}}{=} (\theta_1,\ldots,\theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$.

Soit les probabilités vers l'avant

$$\vec{\alpha}_1(j) \overset{\text{def}}{=} \mathbb{P}[X_1, C_1 = j] = \mathbb{P}[X_1 \mid C_1 = j]\mathbb{P}[C_1 = j] = \mathfrak{u}_j(1) \cdot p_j(x_1), \quad \forall\, j \in \mathcal{C} \tag{II.3.25}$$

et

$$\vec{\alpha}_t(j) \overset{\text{def}}{=} \mathbb{P}[X_{1:t}, C_t = j], \tag{II.3.26}$$

$$= \left(\sum_{i=1}^{K} \vec{\alpha}_{t-1}(i)\Upsilon_{ij}(t-1)\right) p_j(x_t), \quad \begin{cases} \forall\, j \in \mathcal{C} \text{ et} \\ \forall\, t \in \{2,\ldots,T\}, \end{cases} \tag{II.3.27}$$

ainsi que la formule pour la vraisemblance

$$\mathcal{L}_T(\boldsymbol{\theta}) = \sum_{j=1}^{K} \vec{\alpha}_T(j), \tag{II.3.28}$$

présentées dans le cadre de l'algorithme avant à l'appendice II.1.

Soit les probabilités

$$\vec{a}_1(j) \overset{\text{def}}{=} \mathbb{P}[X_1, C_1 = j] = \vec{\alpha}_1(j), \quad \forall\, j \in \mathcal{C}, \tag{II.3.29}$$

$$\vec{a}_t(j) \overset{\text{def}}{=} \mathbb{P}[X_t, C_t = j \mid X_{1:t-1}],$$

$$= \frac{\mathbb{P}[X_{1:t}, C_t = j]}{\mathbb{P}[X_{1:t-1}]} = \frac{\vec{\alpha}_t(i)}{\mathbb{P}[X_{1:t-1}]}, \quad \begin{cases} \forall\, j \in \mathcal{C} \text{ et} \\ \forall\, t \in \{2,\ldots,T\}, \end{cases} \tag{II.3.30}$$

et

$$a_t \overset{\text{def}}{=} \sum_{j=1}^{K} \vec{a}_t(j) = \mathbb{P}[X_t \mid X_{1:t-1}], \quad \forall\, t \in \{1,\ldots,T\}, \tag{II.3.31}$$

obtenues grâce au filtre d'Hamilton présenté à la section 3.6.1.1.

Algorithme 6 : Calcul récursif du score

Soit la notation suivante pour la dérivée terme par terme d'une matrice quelconque $\mathbf{A}_{(K\times K)}$:

$$\left[\frac{\partial \mathbf{A}}{\partial \theta_d}\right] \stackrel{\text{def}}{=} \left(\frac{\partial A_{ij}}{\partial \theta_d}\right)_{i,j\in\{1,\dots,K\}} = \begin{bmatrix} \frac{\partial A_{11}}{\partial \theta_d} & \dots & \frac{\partial A_{1K}}{\partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial A_{K1}}{\partial \theta_d} & \dots & \frac{\partial A_{KK}}{\partial \theta_d} \end{bmatrix}, \tag{II.3.21}$$

qui s'étend naturellement à la dérivée d'un vecteur.
Soit les notations suivante, et leurs interprétation probabilistes :

$$\vec{a}_t \stackrel{\text{def}}{=} (\mathbb{P}[X_t, C_t = 1 \mid X_{1:t-1}], \dots, \mathbb{P}[X_t, C_t = K \mid X_{1:t-1}]), \tag{II.3.22}$$

$$a_t \stackrel{\text{def}}{=} \mathbb{P}[X_t \mid X_{1:t-1}], \tag{II.3.23}$$

$$\vec{\omega}_t \stackrel{\text{def}}{=} (\mathbb{P}[C_t = 1 \mid X_{1:t}], \dots, \mathbb{P}[C_t = K \mid X_{1:t}]). \tag{II.3.24}$$

Soit $\odot$ le produit terme par terme de Hadamard.

Étape 1

$$\vec{\psi}_1(\theta_d) \quad = \quad \left[\frac{\partial \vec{\mathfrak{u}}}{\partial \theta_d}(1)\right] \odot \vec{p}(x_1) \; + \; \vec{\mathfrak{u}}(1) \odot \left[\frac{\partial \vec{p}}{\partial \theta_d}(x_1)\right]$$

Étape 2

Pour $t = 2, \dots, T$, calculer les termes :

$$\vec{\psi}_t(\theta_d) = \left(\frac{\vec{\psi}_{t-1}(\theta_d)}{a_{t-1}} \mathbf{\Gamma}_{t-1} + \vec{\omega}_{t-1}\left[\frac{\partial \mathbf{\Gamma}_{t-1}}{\partial \theta_d}\right]\right) \odot \vec{p}(x_t) + (\vec{\omega}_{t-1}\mathbf{\Gamma}_{t-1}) \odot \left[\frac{\partial \vec{p}}{\partial \theta_d}(x_t)\right]$$

Étape 3

$$\mathcal{S}(\theta_d \mid x_{1:T}) \quad = \quad \vec{\psi}_T(\theta_d)\mathbf{1}_K$$

Coût computationnel

$\mathrm{O}(DTK^2)$

Référence

Version initiale présentée par Lystig et Hughes [2002]

Remarque

Les notations présentées ci-haut proviennent du filtre d'Hamilton (Algorithme 1). Cette algorithme sous forme vectorielle est une adaptation de l'algorithme présentée à la proposition II.3.1.

Soit le score en θ_d, $d \in \{1, \dots, D\}$,

$$\boldsymbol{\mathcal{S}}(\,\theta_d \mid x_{1:T}\,) \stackrel{\text{def}}{=} \frac{\partial \ell_T}{\partial \theta_d}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_d}\log(\mathbb{P}[X_{1:T}]) = \frac{\frac{\partial}{\partial \theta_d}\mathbb{P}[X_{1:T}]}{\mathbb{P}[X_{1:T}]}, \tag{II.3.32}$$

$$= \frac{\frac{\partial}{\partial \theta_d}\sum\limits_{j=1}^{K}\vec{\boldsymbol{\alpha}}_T(j)}{\mathbb{P}[X_{1:T}]}, \tag{II.3.33}$$

$$= \frac{\sum\limits_{j=1}^{K}\left[\frac{\partial}{\partial \theta_d}\vec{\boldsymbol{\alpha}}_T(j)\right]}{\mathbb{P}[X_{1:T}]} = \star. \tag{II.3.34}$$

Obtenir une expression analytique pour le score en θ_d n'est alors qu'une application de la règle du produit et de la règle de dérivation en chaîne (ou règle de dérivation des formules composées). Ce n'est cependant pas exactement cette expression que nous recherchons. En effet, puisque nous savons que les probabilités vers l'avant sont sensibles au problème de soupassement arithmétique, Lystig et Hughes [2002] proposent la séparation par le théorème de Bayes de la probabilité $\mathbb{P}[X_{1:T}] = \mathbb{P}[X_{1:T-1}]\mathbb{P}[\,X_T \mid X_{1:T-1}\,]$ et l'inclusion du terme $\mathbb{P}[X_{1:T-1}]$ dans la somme au numérateur. Cette manoeuvre est ensuite répétée à chaque pas de la dérivation en chaîne afin de mettre à l'échelle chacun des termes de la somme imbriquée résultante. Ainsi,

$$\star = \frac{\sum_{j=1}^{K} \frac{\left[\frac{\partial}{\partial\theta_d}\vec{\alpha}_T(j)\right]}{\mathbb{P}[X_{1:T-1}]}}{\mathbb{P}[X_T \mid X_{1:T-1}]}, \tag{II.3.35}$$

$$= \frac{\sum_{j=1}^{K} \frac{\frac{\partial}{\partial\theta_d}\left(\sum_{i=1}^{K} \vec{\alpha}_{T-1}(i)\gamma_{ij}(T-1)p_j(x_T)\right)}{\mathbb{P}[X_{1:T-1}]}}{\mathbb{P}[X_T \mid X_{1:T-1}]}, \tag{II.3.36}$$

$$= \frac{\sum_{j=1}^{K} \frac{\sum_{i=1}^{K} \frac{\partial}{\partial\theta_d}\left\{\vec{\alpha}_{T-1}(i)\cdot\gamma_{ij}(T-1)\cdot p_j(x_T)\right\}}{\mathbb{P}[X_{1:T-1}]}}{\mathbb{P}[X_T \mid X_{1:T-1}]}, \tag{II.3.37}$$

$$= \frac{\sum_{j=1}^{K} \frac{\sum_{i=1}^{K} \left\{\left[\frac{\partial}{\partial\theta_d}\vec{\alpha}_{T-1}(i)\right]\gamma_{ij}(T-1)\cdot p_j(x_T) + \vec{\alpha}_{T-1}(i)\left[\frac{\partial}{\partial\theta_d}\gamma_{ij}(T-1)\right]p_j(x_T) + \vec{\alpha}_{T-1}(i)\cdot\gamma_{ij}(T-1)\left[\frac{\partial}{\partial\theta_d}p_j(x_T)\right]\right\}}{\mathbb{P}[X_{T-1} \mid X_{1:T-2}]\cdot\mathbb{P}[X_{1:T-2}]}}{\mathbb{P}[X_T \mid X_{1:T-1}]}, \tag{II.3.38}$$

$$= \frac{\sum_{j=1}^{K}\left[\sum_{i=1}^{K}\left\{\begin{array}{l}\frac{\left[\frac{\partial}{\partial\theta_d}\vec{\alpha}_{T-1}(i)\right]}{\mathbb{P}[X_{1:T-2}]}\gamma_{ij}(T-1)\cdot p_j(x_T) \\ +\frac{\vec{\alpha}_{T-1}(i)}{\mathbb{P}[X_{1:T-2}]}\left[\frac{\partial}{\partial\theta_d}\gamma_{ij}(T-1)\right]p_j(x_T) \\ +\frac{\vec{\alpha}_{T-1}(i)}{\mathbb{P}[X_{1:T-2}]}\gamma_{ij}(T-1)\left[\frac{\partial}{\partial\theta_d}p_j(x_T)\right]\end{array}\right\}\right] \times (\mathbb{P}[X_{T-1} \mid X_{1:T-2}])^{-1}}{\mathbb{P}[X_T \mid X_{1:T-1}]}. \tag{II.3.39}$$

Nous reconnaissons dès lors le lien entre les termes en surbrillance. En notant

$$\psi_t(i;\theta_d) \stackrel{\text{def}}{=} \frac{\frac{\partial}{\partial\theta_d}\mathbb{P}[X_{1:t}, C_t = i]}{\mathbb{P}[X_{1:t-1}]}, \tag{II.3.40}$$

$$= \frac{\frac{\partial}{\partial\theta_d}\vec{\alpha}_t(i)}{\mathbb{P}[X_{1:t-1}]} \quad \begin{cases}\forall\, i \in \{1,\dots,K\}, \\ \forall\, d \in \{1,\dots,D\} \text{ et} \\ \forall\, t \in \{1,\dots,T\},\end{cases} \tag{II.3.41}$$

et avec $\vec{a}_{T-1}(j) = \frac{\vec{\alpha}_{T-1}(i)}{\mathbb{P}[X_{1:T-2}]}$, nous avons que

$$\star = \frac{\sum_{j=1}^{K} \psi_T(j;\theta_d)}{\mathbb{P}[X_T \mid X_{1:T-1}]}, \tag{II.3.42}$$

$$= \frac{\sum_{j=1}^{K} \left[\sum_{i=1}^{K} \left\{ \begin{array}{l} \psi_{T-1}(i;\theta_d)\, \gamma_{ij}(T-1) \cdot p_j(x_T) \\ +\vec{a}_{T-1}(j)\left[\frac{\partial}{\partial\theta_d}\gamma_{ij}(T-1)\right] p_j(x_T) \\ +\vec{a}_{T-1}(j)\gamma_{ij}(T-1)\left[\frac{\partial}{\partial\theta_d} p_j(x_T)\right] \end{array} \right\} \times \left(\mathbb{P}[X_{T-1} \mid X_{1:T-2}]\right)^{-1} \right]}{\mathbb{P}[X_T \mid X_{1:T-1}]} \tag{II.3.43}$$

et donc que

$$\begin{aligned} \psi_T(j;\theta_d) = \sum_{i=1}^{K} \Bigg\{ & \psi_{T-1}(i;\theta_d)\, \gamma_{ij}(T-1) \cdot p_j(x_T) \\ & + \vec{a}_{T-1}(j)\left[\frac{\partial}{\partial\theta_d}\gamma_{ij}(T-1)\right] p_j(x_T) \\ & + \vec{a}_{T-1}(j)\gamma_{ij}(T-1)\left[\frac{\partial}{\partial\theta_d} p_j(x_T)\right] \Bigg\} \times (a_{T-1})^{-1}, \quad \forall\, d \in \{1,\ldots,D\}. \end{aligned} \tag{II.3.44}$$

Cependant, puisqu'une probabilité par ligne de chaque matrice de transition $\mathbf{\Gamma}_t$ peut être obtenue par soustraction, ce résultat n'est valide que pour $j = 1,\ldots,K-1$. En effet, puisque chaque matrice de transition n'a que $K(K-1)$ paramètres libres, il faut adapter la relation (II.3.44) pour le cas $j = K$.

Comme à la section 1.3 sur les modèles de mélange, nous choisissons, par simplicité et sans perte de généralité, d'exprimer les probabilités de transition vers le régime K en fonction des autres probabilités sur la même ligne, soit

$$\gamma_{iK}(t) = 1 - \sum_{j=1}^{K-1} \gamma_{ij}(t), \quad \begin{cases} \forall\, i \in \{1,\ldots,K\} \text{ et} \\ \forall\, t \in \{1,\ldots,T\}. \end{cases} \tag{II.3.45}$$

Ainsi, nous avons en $j = K$ que

$$\psi_T(K;\theta_d) \stackrel{\text{def}}{=} \frac{\frac{\partial}{\partial\theta_d}\mathbb{P}[X_{1:T}, C_T = K]}{\mathbb{P}[X_{T-1} \mid X_{1:T-2}]}, \tag{II.3.46}$$

$$\begin{aligned}
&= \sum_{i=1}^{K} \Bigg\{ \psi_{T-1}(i;\theta_d)\left(1 - \sum_{j=1}^{K-1} \gimel_{ij}(T-1)\right) p_K(x_T) \\
&\qquad + \vec{a}_{T-1}(i)\left[\frac{\partial}{\partial\theta_d}\left(1 - \sum_{j=1}^{K-1} \gimel_{ij}(T-1)\right)\right] p_K(x_T) \\
&\qquad + \vec{a}_{T-1}(i)\left(1 - \sum_{j=1}^{K-1} \gimel_{ij}(T-1)\right)\left[\frac{\partial p_K}{\partial\theta_d}(x_T)\right] \Bigg\} \times (a_{T-1})^{-1},
\end{aligned} \tag{II.3.47}$$

$$\begin{aligned}
&= \sum_{i=1}^{K} \Bigg\{ \psi_{T-1}(i;\theta_d) \cdot \gimel_{iK}(T-1) \cdot p_K(x_T) \\
&\qquad - \vec{a}_{T-1}(i)\left(\sum_{j=1}^{K-1}\left[\frac{\partial \gimel_{ij}}{\partial\theta_d}(T-1)\right]\right) p_K(x_T) \\
&\qquad + \vec{a}_{T-1}(i) \cdot \gimel_{iK}(T-1) \cdot \left[\frac{\partial p_K}{\partial\theta_d}(x_T)\right] \Bigg\} \times (a_{T-1})^{-1}, \quad \forall\, d \in \{1,\ldots,D\}.
\end{aligned} \tag{II.3.48}$$

Puisque la relation (II.3.27) (p.153) utilisée à l'équation (II.3.37) (p.156) n'est valide que pour $t \in \{2,\ldots,T\}$, les relations (II.3.44) (p.157) et (II.3.48) ne tiennent que sur cet ensemble. Il faut donc obtenir séparément une expression pour $\psi_1(i;\theta_d)$, $\forall\, j \in 1,\ldots,K$. Comme précédemment, il faut séparer les cas $j = 1,\ldots,K-1$ et $j = K$. Nous avons donc finalement que

$$\begin{aligned}
\psi_1(j;\theta_d) &\stackrel{\text{def}}{=} \frac{\partial}{\partial\theta_d}\mathbb{P}[X_1, C_1 = j], \\
&= \frac{\partial}{\partial\theta_d}(\mathbb{P}[C_t = j] \cdot \mathbb{P}[X_1 \mid C_1 = j\,]), \\
&= \left[\frac{\partial \mathfrak{u}_j}{\partial\theta_d}(1)\right] p_j(x_1) + \mathfrak{u}_j(1)\left[\frac{\partial p_j}{\partial\theta_d}(x_1)\right], \quad \begin{cases} \forall\, j \in \{1,\ldots,K-1\} \text{ et} \\ \forall\, d \in \{1,\ldots,D\} \end{cases}
\end{aligned} \tag{II.3.49}$$

et

$$\psi_1(K;\theta_d) = \left(-\sum_{i=1}^{K-1}\left[\frac{\partial \mathfrak{u}_i}{\partial\theta_d}(1)\right]\right) p_j(x_1) + \left(1 - \sum_{i=1}^{K-1} \mathfrak{u}_i(1)\right)\left[\frac{\partial p_j}{\partial\theta_d}(x_1)\right], \quad \forall\, d \in \{1,\ldots,D\}. \tag{II.3.50}$$

□

II.3.4.2. Algorithme récursif pour la matrice hessienne de $\ell_T(\boldsymbol{\theta})$

Pour ce qui est maintenant de construire la matrice hessienne, la proposition II.3.2 présente la méthode permettant de la construire récursivement. Cependant, bien que cet algorithme soit intéressant d'un point de vue théorique, celui-ci nécessite $\mathrm{O}(D^2TK^2)$ opérations, ce qui limite grandement sont champ d'applications. Puisque cette matrice nous servirait principalement à estimer la matrice d'information de Fisher qui sera présentée à l'appendice II.3.6, nous y reportons la discussion concernant les méthodes permettant de contourner ce problème.

Proposition II.3.2 (Algorithme récursif pour la matrice hessienne de $\ell_T(\boldsymbol{\theta})$)**.** *Soit le vecteur de paramètres libres du modèle* $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\theta_1, \ldots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$.

Soit les dérivées partielles mises à l'échelle de $\mathbb{P}[X_{1:t}, C_t = j]$ *obtenues à la proposition II.3.1 et notées*

$$\psi_t(j;\theta_d) \stackrel{\text{def}}{=} \frac{\frac{\partial}{\partial\theta_d}\mathbb{P}[X_{1:t}, C_t = j]}{\mathbb{P}[X_{1:t-1}]} = \frac{\frac{\partial}{\partial\theta_d}\vec{\boldsymbol{\alpha}}_t(j)}{\mathbb{P}[X_{1:t-1}]}, \quad \begin{cases} \forall\, j \in \{1, \ldots, K\}, \\ \forall\, d \in \{1, \ldots, D\} \text{ et} \\ \forall\, t \in \{1, \ldots, T\}, \end{cases} \tag{II.3.51}$$

avec les probabilité vers l'avant

$$\vec{\boldsymbol{\alpha}}_t(j) \stackrel{\text{def}}{=} \mathbb{P}[X_{1:t}, C_t = j], \tag{II.3.52}$$

$$= \sum_{i=1}^{K} \vec{\boldsymbol{\alpha}}_{t-1}(i) \cdot \Upsilon_{ij}(t-1) \cdot p_j(x_t), \quad \forall\, j \in \mathcal{C}, \tag{II.3.53}$$

provenant de l'algorithme avant.

Soit la notation pour les dérivées partielles de second ordre de $\mathbb{P}[X_{1:t}, C_t = j]$ *mises à l'échelle*

$$\phi_t(j;\theta_d,\theta_e) \stackrel{\text{def}}{=} \frac{\frac{\partial^2}{\partial\theta_d\partial\theta_e}\mathbb{P}[X_{1:t}, C_t = j]}{\mathbb{P}[X_{1:t-1}]}, \quad \begin{cases} \forall\, j \in \{1, \ldots, K\}, \\ \forall\, d,e \in \{1, \ldots, D\} \text{ et} \\ \forall\, t \in \{1, \ldots, T\}, \end{cases} \tag{II.3.54}$$

Ces dernières peuvent être obtenues par les relations de récurrence

$$\begin{aligned}
\phi_t(j;\theta_d,\theta_e) = \Bigg[\sum_{i=1}^{K} \Bigg\{ & \phi_{t-1}(i;\theta_d,\theta_e)\cdot \gamma_{ij}(t-1)\cdot p_j(x_t) \\
& + \psi_{t-1}(i;\theta_d)\left[\frac{\partial \gamma_{ij}}{\partial\theta_e}(t-1)\right] p_j(x_t) \\
& + \psi_{t-1}(i;\theta_d)\cdot \gamma_{ij}(t-1)\left[\frac{\partial p_j}{\partial\theta_e}(x_t)\right] \\
& + \psi_{t-1}(i;\theta_e)\left[\frac{\partial \gamma_{ij}}{\partial\theta_d}(t-1)\right] p_j(x_t) \\
& + \psi_{t-1}(i;\theta_e)\cdot \gamma_{ij}(t-1)\left[\frac{\partial p_j}{\partial\theta_d}(x_t)\right] \\
& + \vec{a}_{t-1}(i)\left[\frac{\partial^2 \gamma_{ij}}{\partial\theta_d\partial\theta_e}(t-1)\right] p_j(x_t) \\
& + \vec{a}_{t-1}(i)\left[\frac{\partial \gamma_{ij}}{\partial\theta_d}(t-1)\right]\left[\frac{\partial p_j}{\partial\theta_e}(x_t)\right] \\
& + \vec{a}_{t-1}(i)\left[\frac{\partial \gamma_{ij}}{\partial\theta_e}(t-1)\right]\left[\frac{\partial p_j}{\partial\theta_d}(x_t)\right] \\
& + \vec{a}_{t-1}(i)\cdot \gamma_{ij}(t-1)\left[\frac{\partial^2 p_j}{\partial\theta_d\partial\theta_e}(x_t)\right]\Bigg\}\Bigg] \times (a_{t-1})^{-1}, \quad \begin{cases} \forall\, j\in\{1,\dots,K\}, \\ \forall\, d,e\in\{1,\dots,D\}\ et \\ \forall\, t\in\{2,\dots,T\}, \end{cases}
\end{aligned} \tag{II.3.55}$$

et initialisées en $t=1$ par les relations

$$\phi_1(j;\theta_d,\theta_e) \overset{\text{def}}{=} \frac{\partial^2}{\partial\theta_d\partial\theta_e}\mathbb{P}[X_1,C_1=j], \tag{II.3.56}$$

$$\begin{aligned}
= & \left[\frac{\partial^2 \mathfrak{u}_j}{\partial\theta_d\partial\theta_e}(1)\right] p_j(x_1) \\
& + \left[\frac{\partial \mathfrak{u}_j}{\partial\theta_d}(1)\right]\left[\frac{\partial p_j}{\partial\theta_e}(x_1)\right] \\
& + \left[\frac{\partial \mathfrak{u}_j}{\partial\theta_e}(1)\right]\left[\frac{\partial p_j}{\partial\theta_d}(x_1)\right] \\
& + \mathfrak{u}_j(1)\left[\frac{\partial^2 p_j}{\partial\theta_d\partial\theta_e}(x_1)\right], \quad \begin{cases} \forall\, j\in\{1,\dots,K\}\ et \\ \forall\, d,e\in\{1,\dots,D\}, \end{cases}
\end{aligned} \tag{II.3.57}$$

avec à nouveau

$$\left[\frac{\partial \Upsilon_{iK}}{\partial \theta_d}(t-1)\right] = \left(-\sum_{j=1}^{K-1}\left[\frac{\partial \Upsilon_{ij}}{\partial \theta_d}(t-1)\right]\right), \quad \forall\, i \in \{1,\ldots,K\} \tag{II.3.58}$$

et

$$\left[\frac{\partial \mathfrak{u}_K}{\partial \theta_d}(1)\right] = \left(-\sum_{i=1}^{K-1}\left[\frac{\partial \mathfrak{u}_i}{\partial \theta_d}(1)\right]\right), \tag{II.3.59}$$

ainsi que

$$\left[\frac{\partial^2 \Upsilon_{iK}}{\partial \theta_d \partial \theta_e}(t-1)\right] = \left(-\sum_{j=1}^{K-1}\left[\frac{\partial^2 \Upsilon_{ij}}{\partial \theta_d \partial \theta_e}(t-1)\right]\right), \quad \forall\, i \in \{1,\ldots,K\} \tag{II.3.60}$$

et

$$\left[\frac{\partial^2 \mathfrak{u}_K}{\partial \theta_d \partial \theta_e}(1)\right] = \left(-\sum_{i=1}^{K-1}\left[\frac{\partial^2 \mathfrak{u}_j}{\partial \theta_d \partial \theta_e}(1)\right]\right). \tag{II.3.61}$$

Les éléments de la matrice hessienne de $\ell_T(\boldsymbol{\theta})$ sont alors donnés par

$$[\mathfrak{H}_{\ell_T}(\boldsymbol{\theta})]_{de} = \frac{\partial^2 \ell_T}{\partial \theta_d \partial \theta_e}(\boldsymbol{\theta}) = \frac{\partial^2 \ell_T}{\partial \theta_e \partial \theta_d}(\boldsymbol{\theta}) = [\mathfrak{H}_{\ell_T}(\boldsymbol{\theta})]_{ed}, \tag{II.3.62}$$

$$= \frac{\Phi_T(\theta_d, \theta_e)}{a_T} - \frac{\Psi_T(\theta_d)\cdot \Psi_T(\theta_e)}{(a_T)^2}, \quad \forall\, d, e \in \{1,\ldots,D\} \tag{II.3.63}$$

avec $\Psi_T(\theta_d) = \sum_{i=1}^{K} \psi_T(i;\theta_d)$ *et* $\Phi_t(\theta_d,\theta_e) = \sum_{j=1}^{K} \phi_t(j;\theta_d,\theta_e)$.

Par le fait même, nous obtenons les termes permettant de construire la matrice d'information observée, correspondant au négatif de la matrice hessienne de la log-vraisemblance, soit

$$[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta} \mid x_{1:T})]_{de} = -[\mathfrak{H}_{\ell_T}(\boldsymbol{\theta})]_{de}, \quad \forall\, d, e \in \{1,\ldots,D\}. \tag{II.3.64}$$

Démonstration. La preuve de cette proposition fait principalement usage de la règle de dérivation d'un produit, de la règle de dérivation en chaîne et du théorème de Bayes, tout comme la proposition II.3.1. Ainsi, la preuve de cette proposition étant très semblable à celle de la proposition II.3.1 et surtout très lourde en notation, nous choisissons d'omettre celle-ci et de diriger les intéressé.e.s vers Lystig et Hughes [2002](p. 688). □

II.3.4.3. Enjeux liés à la dimension du vecteur $\boldsymbol{\theta}$

Nous avons jusqu'à maintenant défini $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^D$, sans nous intéresser réellement à la valeur que prend D dans nos modèles. C'est pourtant une quantité très importante apparaissant entre autres dans le calcul du nombre d'opérations requises pour les algorithmes présentés dans la présente section.

Bien entendu, le cas où la matrice de transition n'est pas homogène et ne s'est vue imposer aucune structure particulière (i.e. que les matrices de la suite $\{\mathbf{\Gamma}_t\}_{t\in\{1,\dots,T\}}$ nécessitent chacunes $K(K-1)$ paramètres libres) peut être exclus d'emblée. En effet, cette structure ne permettant ni de faire de l'inférence sur le processus latent, ni de faire des prévisions ou d'étendre le modèle au dela de $t = T$, elle n'a aucun intérêt ormis théorique. Ainsi, à l'instar de l'inférence en $t + h$ conditionnellement à $\boldsymbol{\mathcal{F}}_t$ présentée à la section 3.6.1.2, nous nous intéresserons ici au cas où la matrice de transition est homogène, mais où la distribution initiale $\vec{\mathfrak{u}}(1)$ n'est pas spécifiée et doit donc être estimée. Nous réservons pour le chapitre 5 les discussions et résultats pour des architectures spécifiques de $\mathbf{\Gamma}_t$

Puisque nous construisons notre modèle autour de distributions normales univariées $(X_t \mid C_t = i \sim \mathcal{N}(\mu_i,\sigma_i^2), \quad \forall\, i \in \{1,\dots,K\})$ nécessitant chacune un paramètre de position μ_i et un paramètre d'échelle σ_i, nous aurons d'abord besoin de $2K$ paramètres afin de paramétrer ces K distributions normales.

En ce qui a trait à la chaîne de Markov, nous avons déjà indiqué précédemment que la construction d'une matrice de transition homogène nécessite $K(K-1)$ paramètres libres, avec une croissance quadratique en fonction du nombre de régimes du modèle. Si de plus la distribution initiale de la chaîne de Markov n'est pas spécifiée dans la définition du modèle, il faudra retenir $(K-1)$ paramètres pour celle-ci.

Cette spécification implique donc un total de $D = 2K + K(K-1) + (K-1) = \mathrm{O}(K^2)$ paramètres libres. Nous savons que l'algorithme permettant de calculer le score (Algorithme algorithme 6 : Calcul récursif du score (p.154)) nécessite $\mathrm{O}(DTK^2)$ opérations et que celui pour obtenir la matrice hessienne nécessite $\mathrm{O}(D^2TK^2)$ opérations. Dans le modèle considéré aux paragraphes précédents, ceci correspond à une complexité de $\mathrm{O}(TK^4)$ pour le score et de $\mathrm{O}(TK^6)$ pour la matrice hessienne, en plus d'un espace machine considérable.

Nous présentons ici, à titre d'illustration, la forme que prend le score lorsque la chaîne de Markov est homogène. La démarche pour le calcul de la matrice hessienne est omise car sa complexité est démesurée et qu'elle ressemble énormément à celle pour le score.

Exemple II.3.1 (Calcul du score d'un *HMM* homogène à $K = 3$ régimes)**.** *Soit un modèle de Markov à variables latentes homogène à* 3 *régimes et à composantes normales univariées de densités*

$$p_i(x_t) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2\,\sigma_i}(x-\mu_i)^2\right\}, \quad \forall\, i \in \mathcal{C}. \tag{II.3.65}$$

Ce modèle est complètement défini par les vecteurs de paramètres

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)', \tag{II.3.66}$$

$$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)' \tag{II.3.67}$$

et

$$\vec{\mathfrak{u}}(1) = (\mathfrak{u}_1(1), \ldots, \mathfrak{u}_K(1)), \tag{II.3.68}$$

ainsi que la matrice de transition unique

$$\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j\in C}. \tag{II.3.69}$$

Le vecteur de paramètre de notre modèle est donc

$$\boldsymbol{\theta} = (\mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K, \mathfrak{u}_1(1), \ldots, \mathfrak{u}_{K-1}(1), \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32})' \tag{II.3.70}$$

et en nous fiant à la proposition II.3.1, nous aurons besoin de calculer, $\forall\, \theta_d \in \boldsymbol{\theta}$ *et* $\forall\, t \in \{1, \ldots, T\}$*, les dérivées*

$$\frac{\partial \mathfrak{u}_1}{\partial \theta_d}(1),\ \frac{\partial \mathfrak{u}_2}{\partial \theta_d}(1),\ \frac{\partial \gamma_{11}}{\partial \theta_d},\ \frac{\partial \gamma_{12}}{\partial \theta_d},\ \frac{\partial \gamma_{21}}{\partial \theta_d},\ \frac{\partial \gamma_{22}}{\partial \theta_d},\ \frac{\partial \gamma_{31}}{\partial \theta_d},\ \frac{\partial \gamma_{32}}{\partial \theta_d},\ \frac{\partial p_1}{\partial \theta_d}(x_t),\ \frac{\partial p_2}{\partial \theta_d}(x_t) \text{ et } \frac{\partial p_3}{\partial \theta_d}(x_t). \tag{II.3.71}$$

Ces dérivées partielles sont rassemblées dans le tableau II.1.

Tableau II.1 – Dérivées partielles en fonction des paramètres naturels libres

	$\mathfrak{u}_1(1)$	$\mathfrak{u}_2(1)$	$\mathfrak{Y}_{11}$	$\mathfrak{Y}_{12}$	$\mathfrak{Y}_{21}$	$\mathfrak{Y}_{22}$	$\mathfrak{Y}_{31}$	$\mathfrak{Y}_{32}$	$p_1(x_t)$	$p_2(x_t)$	$p_3(x_t)$
μ_1	0	0	0	0	0	0	0	0	$\left(\frac{x_t-\mu_1}{\sigma_1}\right)\cdot p_1(x_t)$	0	0
μ_2	0	0	0	0	0	0	0	0	0	$\left(\frac{x_t-\mu_2}{\sigma_2}\right)\cdot p_2(x_t)$	0
μ_3	0	0	0	0	0	0	0	0	0	0	$\left(\frac{x_t-\mu_3}{\sigma_3}\right)\cdot p_3(x_t)$
σ_1	0	0	0	0	0	0	0	0	$\frac{1}{\sigma_1}\left(\frac{\left(x_t-\mu_1\right)^2}{2\,\sigma_1}-1\right)\cdot p_1(x_t)$	0	0
σ_2	0	0	0	0	0	0	0	0	0	$\frac{1}{\sigma_2}\left(\frac{\left(x_t-\mu_2\right)^2}{2\,\sigma_2}-1\right)\cdot p_2(x_t)$	0
σ_3	0	0	0	0	0	0	0	0	0	0	$\frac{1}{\sigma_3}\left(\frac{\left(x_t-\mu_3\right)^2}{2\,\sigma_3}-1\right)\cdot p_3(x_t)$
$\mathfrak{u}_1(1)$	1	0	0	0	0	0	0	0	0	0	0
$\mathfrak{u}_2(1)$	0	1	0	0	0	0	0	0	0	0	0
$\mathfrak{Y}_{11}$	0	0	1	0	0	0	0	0	0	0	0
$\mathfrak{Y}_{12}$	0	0	0	1	0	0	0	0	0	0	0
$\mathfrak{Y}_{21}$	0	0	0	0	1	0	0	0	0	0	0
$\mathfrak{Y}_{22}$	0	0	0	0	0	1	0	0	0	0	0
$\mathfrak{Y}_{31}$	0	0	0	0	0	0	1	0	0	0	0
$\mathfrak{Y}_{32}$	0	0	0	0	0	0	0	1	0	0	0

Avec ces dérivées partielles en main, nous avons

$$\left[\frac{\partial\vec{\mathfrak{u}}}{\partial\theta_d}(1)\right] \stackrel{\text{def}}{=} \left(\frac{\partial\mathfrak{u}_1}{\partial\theta_d}(1),\ldots,\frac{\partial\mathfrak{u}_K}{\partial\theta_d}(1)\right)\Big|_{K=3} = \left(\frac{\partial\mathfrak{u}_1}{\partial\theta_d}(1),\ldots,\frac{\partial\mathfrak{u}_{K-1}}{\partial\theta_d}(1),\left(-\sum_{i=1}^{K-1}\frac{\partial\mathfrak{u}_i}{\partial\theta_d}(1)\right)\right)\Big|_{K=3},$$

$$= \begin{cases} (1,0,-1), & si\ \theta_d = \mathfrak{u}_1(1), \\ (0,1,-1), & si\ \theta_d = \mathfrak{u}_1(2), \\ (0,0,0), & sinon, \end{cases} \tag{II.3.72}$$

$$\left[\frac{\partial\vec{p}}{\partial\theta_d}(x_t)\right] \stackrel{\text{def}}{=} \left(\frac{\partial p_1}{\partial\theta_d}(x_t),\ldots,\frac{\partial p_K}{\partial\theta_d}(x_t)\right)\Big|_{K=3},$$

$$= \begin{cases} \left(\left(\frac{x_t-\mu_1}{\sigma_1}\right)\cdot p_1(x_t),\quad 0,\quad 0\right), & si\ \theta_d = \mu_1\,, \\ \left(0,\quad \left(\frac{x_t-\mu_2}{\sigma_2}\right)\cdot p_2(x_t),\quad 0\right), & si\ \theta_d = \mu_2\,, \\ \left(0,\quad 0,\quad \left(\frac{x_t-\mu_3}{\sigma_3}\right)\cdot p_3(x_t)\right), & si\ \theta_d = \mu_3\,, \\ \left(\frac{1}{\sigma_1}\left(\frac{\left(x_t-\mu_1\right)^2}{2\,\sigma_1}-1\right)\cdot p_1(x_t),\quad 0,\quad 0\right), & si\ \theta_d = \sigma_1\,, \\ \left(0,\quad \frac{1}{\sigma_2}\left(\frac{\left(x_t-\mu_2\right)^2}{2\,\sigma_2}-1\right)\cdot p_2(x_t),\quad 0\right), & si\ \theta_d = \sigma_2\,, \\ \left(0,\quad 0,\quad \frac{1}{\sigma_3}\left(\frac{\left(x_t-\mu_3\right)^2}{2\,\sigma_3}-1\right)\cdot p_3(x_t)\right), & si\ \theta_d = \sigma_3\,, \\ \left(0,\quad 0,\quad 0\right), & sinonx \end{cases} \tag{II.3.73}$$

et

$$\left[\frac{\partial\boldsymbol{\Gamma}}{\partial\theta_d}\right] \stackrel{\text{def}}{=} \left(\frac{\partial\gamma_{ij}}{\partial\theta_d}\right)_{i,j\in\mathcal{C}} = \left[\begin{array}{ccc} \frac{\partial\gamma_{11}}{\partial\theta_d} & \cdots & \frac{\partial\gamma_{1K}}{\partial\theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial\gamma_{K1}}{\partial\theta_d} & \cdots & \frac{\partial\gamma_{KK}}{\partial\theta_d} \end{array}\right]\Bigg|_{K=3}, \tag{II.3.74}$$

$$= \left[\begin{array}{ccc} \frac{\partial\gamma_{11}}{\partial\theta_d} & \cdots & \frac{\partial\gamma_{1(K-1)}}{\partial\theta_d} \quad \left(-\sum_{j=1}^{K-1}\frac{\partial\gamma_{1j}}{\partial\theta_d}\right) \\ \vdots & \ddots & \vdots \\ \frac{\partial\gamma_{K1}}{\partial\theta_d} & \cdots & \frac{\partial\gamma_{K(K-1)}}{\partial\theta_d} \quad \left(-\sum_{j=1}^{K-1}\frac{\partial\gamma_{Kj}}{\partial\theta_d}\right) \end{array}\right]\Bigg|_{K=3}. \tag{II.3.75}$$

Cette dernière matrice a toujours sensiblement la même forme. Si $\theta_d = \gamma_{ij}$, *alors* $\left[\frac{\partial \mathbf{\Gamma}}{\partial \gamma_{ij}}\right]_{ij} = 1$, $\left[\frac{\partial \mathbf{\Gamma}}{\partial \gamma_{ij}}\right]_{iK} = -1$ *et toutes les autres entrées sont nulles. Pour tous les autres paramètres du modèle, cette matrice est complètement nulle. À titre d'illustration, nous avons par exemple que*

$$\left[\frac{\partial \mathbf{\Gamma}}{\partial \gamma_{11}}\right] = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{II.3.76}$$

II.3.5. Reparamétrisation

En introduction à cette section, il était mentionné que les paramètres de notre modèles étaient pour le moment sous contraintes. En effet, nous avons jusqu'à maintenant imposé les contraintes suivantes sur nos paramètres :

Contrainte 1. $0 \leq \boldsymbol{\sigma} < \infty$,

Contrainte 2. $0 \leq \vec{\mathfrak{u}}(1) \leq 1$,

Contrainte 3. $\sum_{i=1}^{K} \mathfrak{u}_i(1) = 1$,

Contrainte 4. $0 \leq \mathbf{\Gamma}_t \leq 1, \quad \forall\, t \in \mathbb{N}^\star$,

Contrainte 5. $\sum_{j=1}^{K} \gamma_{ij} = 1, \quad \forall\, i \in \mathcal{C}$ et $\forall\, t \in \mathbb{N}^\star$.

Quoique certains outils informatiques mentionnés dans Zucchini et Macdonald [2009](p. 47) permettent d'optimiser une fonction de plusieurs variables sous contraintes et qu'il puisse être contre-indiqué de modifier des variables n'étant contraintes que linéairement, une optimisation non-contrainte savère habituellement beaucoup plus longue et pourrait ne pas converger vers l'EMV. Nous préconiserons donc une reparamétrisation de toutes les variables étant contraintes dans notre modèle. Puisque celles-ci sont claires et concises, nous adoptons une version légèrement modifiée des méthodes de reparamétrisation proposées par Zucchini et Macdonald [2009](p. 48-49).

Remarque II.3.5 (Paramètres de travail). Les paramètres non-contraints ainsi obtenus sont communément appelés les **paramètres de travail** (*Working parameters*), alors que les paramètres initiaux contraints du modèle sont appelés **paramètres naturels** (*Natural parameters*).

Reparamétriser correctement un modèle nécessite de trouver une transformation surjective $g : \mathbb{R}^n \to \boldsymbol{\Theta} \subset \mathbb{R}^D$, avec n le nombre de paramètres de travail et $\boldsymbol{\Theta}$ le domaine de dimension D des paramètres naturels. Il est important que la transformation soit au moins surjective afin de nous assurer de chercher dans tout l'ensemble de départ $\boldsymbol{\Theta}$. Puisque nous notons $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^D$ le vecteur de paramètres naturels, nous adopterons la notation $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^n$ afin d'identifier le vecteur contenant les paramètres de travail.

Remarque II.3.6 (Différentiabilité de la transformation g). Aux fins d'optimisation de la log-vraisemblance d'un modèle de Markov à variables latentes, il sera important que la transformation g choisie soit de surcroît différentiable en $\tilde{\boldsymbol{\theta}}$. Ceci sera évident lorsque nous chercherons à exprimer le score et la matrice hessienne par rapport aux paramétres de travail en fin de section.

Puisque nous voulons maximiser la vraisemblance sans contraintes, nous devrons exprimer celle-ci en fonction des paramètres de travail et obtenir l'EMV correspondant à ceux-ci, que nous noterons $\check{\boldsymbol{\theta}} \in \mathbb{R}^n$. Une propriété importante à cet égard est la propriété d'invariance de l'estimateur du maximum de vraisemblance.

Théorème II.3.1 (Propriété d'invariance de l'EMV)**.** *Si $\hat{\theta}$ est l'EMV de θ, alors pour toute fonction f, $f\big(\hat{\theta}\big)$ est l'EMV de $f(\theta)$.*

Démonstration. Voir le théorème 7.2.10 de Casella et Berger [2001](p. 320). □

En effet, grâce à cette propriété d'invariance et au fait que g est surjective, nous avons que

$$\max_{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^n} \log\Big(\mathbb{P}\Big[\, x_{1:T} \;\Big|\; \boldsymbol{\theta} = g\big(\tilde{\boldsymbol{\theta}}\big) \,\Big]\Big) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log(\mathbb{P}[\, x_{1:T} \mid \boldsymbol{\theta} \,]), \tag{II.3.77}$$

ce qui signifie que si l'EMV $\widehat{\boldsymbol{\theta}}$ est unique, alors nous avons nécessairement que $\widehat{\boldsymbol{\theta}} = g\big(\check{\boldsymbol{\theta}}\big)$. C'est une raison de plus pour laquelle l'identifiabilité du modèle est si importante lors de l'estimation des paramètres.

Une fois la vraisemblance exprimée en fonction de nos paramètres de travail, l'optimisation s'effectue en deux étapes :

Étape 1. Obtenir l'EMV de travail

$$\check{\boldsymbol{\theta}} = \arg\max_{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^n} \log\Big(\mathbb{P}\Big[\, x_{1:T} \;\Big|\; \boldsymbol{\theta} = g\big(\tilde{\boldsymbol{\theta}}\big) \,\Big]\Big). \tag{II.3.78}$$

Étape 2. Utiliser la transformation g afin d'obtenir l'EMV naturel

$$\widehat{\boldsymbol{\theta}} = g\big(\check{\boldsymbol{\theta}}\big). \tag{II.3.79}$$

Sous la spécification généralisée (p.28) d'un *HMM* avec composantes normales, les paramètres à transformer se divise en deux groupes : Les écarts-types liés aux distributions normales, soit $\boldsymbol{\sigma}$, et les paramètres liés au processus de changement de régime, soit les matrices de transition et la distribution initiale si elle n'est pas donnée explicitement dans la définition du modèle—voir l'appendice I.4.

En choisissant d'incorporer les contraintes 3 et 5 dans la démarche des algorithmes récursifs pour le score et la matrice hessienne et d'information observée présentés à la section II.3.4, nous avons dans un premier temps réduit la dimension du vecteur de paramètres naturels de $(K+1)$, mais aussi simplifier la méthode de reparamétrisation à un enjeu de positivité et de normalisation.

II.3.5.1. Méthode pour $\boldsymbol{\sigma}$

Pour les paramètres liés aux distributions de probabilité, la reparamétrisation est relativement simple. Puisque nous cherchons simplement à nous assurer que les paramètres d'écart-type soient non-négatifs, nous aurons usage de la fonction exponentielle et de son inverse, la fonction logarithmique. Si nous définissons

$$\eta_i = \log(\sigma_i), \quad \forall\, i \in \mathcal{C}, \tag{II.3.80}$$

alors $\eta_i \in \mathbb{R}$ et n'est donc plus contraint. La maximisation se fait alors en fonction de cette dernière, puis on applique la transformation inverse

$$\sigma_i = \exp\{\eta_i\}, \tag{II.3.81}$$

qui nous garantie que $\sigma_i \in \mathbb{R}_+$.

II.3.5.2. Méthode pour $\boldsymbol{\Gamma}_t$

Pour ce qui est de reparamétriser la matrice de transition, la procédure est légèrement plus complexe. Nous choisissons par soucis de simplicité d'utiliser à nouveau la fonction exponentielle. Si nous définissons

$$\tau_{ij}(t) = \log\left(\frac{\Upsilon_{ij}(t)}{\Upsilon_{iK}(t)}\right), \begin{cases} \forall\, i \in \{1,\ldots,K\}, \\ \forall\, j \in \{1,\ldots,K-1\} \text{ et} \\ \forall\, t \in \{1,\ldots,T\}, \end{cases} \tag{II.3.82}$$

alors les termes $\tau_{ij}(t) \in \mathbb{R}$ ne sont plus contraints et la transformation inverse

$$\Upsilon_{ij}(t) = \frac{\exp\{\tau_{ij}(t)\}}{1 + \sum_{\ell=1}^{K-1} \exp\{\tau_{i\ell}(t)\}}, \tag{II.3.83}$$

permet de récupérer les probabilités de transition naturelles.

II.3.5.3. Méthode pour $\vec{\mathfrak{u}}(1)$

La même technique que nous venons d'appliquer sur les lignes de $\mathbf{\Gamma}_t$ permet de reparamétriser $\vec{\mathfrak{u}}(1)$. Si nous définissons

$$\upsilon_i = \log\left(\frac{\mathfrak{u}_i(1)}{\mathfrak{u}_K(1)}\right), \forall\, i \in \{1, \ldots, K-1\}, \tag{II.3.84}$$

alors les termes $\upsilon_i \in \mathbb{R}$ ne sont plus contraints et la transformation inverse

$$\mathfrak{u}_i(1) = \frac{\exp\{\upsilon_i\}}{1 + \sum_{\ell=1}^{K-1} \exp\{\upsilon_\ell\}}, \tag{II.3.85}$$

permet de récupérer la distribution initiale naturelle.

Exemple II.3.2 (Reparamétrisation de $\mathbf{\Gamma}_t$ dans le cas $K = 3$)**.** *Soit la matrice de transition au temps $t \in \{1, \ldots, T\}$ d'un* HMM *à $K = 3$ régimes*

$$\mathbf{\Gamma}_t = \begin{bmatrix} \Upsilon_{11}(t) & \Upsilon_{12}(t) & \Upsilon_{13}(t) \\ \Upsilon_{21}(t) & \Upsilon_{22}(t) & \Upsilon_{23}(t) \\ \Upsilon_{31}(t) & \Upsilon_{32}(t) & \Upsilon_{33}(t) \end{bmatrix} \tag{II.3.86}$$

qui, rappelons-le, n'a que $K(K-1) = 6$ paramètres libres.

Soit la matrice

$$\mathbf{T}_t = \begin{bmatrix} \tau_{11}(t) & \tau_{12}(t) & 0 \\ \tau_{21}(t) & \tau_{22}(t) & 0 \\ \tau_{31}(t) & \tau_{32}(t) & 0 \end{bmatrix}, \tag{II.3.87}$$

dont les entrées sont liées aux probabilités de transition par le biais de transformation strictement croissante de l'équation (II.3.82) (p.168) *et où les entrées $\tau_{iK}(t) = 0, \forall\, i \in \{1, \ldots, K\}$.*

Afin de rendre la démarche plus lisible, intéressons-nous à la première ligne de la matrice de transition afin d'illustrer la méthode. Les probabilités sur cette ligne sont telles que

$$\Upsilon_{11}(t) = \frac{\exp\{\tau_{11}(t)\}}{\exp\{\tau_{11}(t)\} + \exp\{\tau_{12}(t)\} + 1}, \tag{II.3.88}$$

$$\Upsilon_{12}(t) = \frac{\exp\{\tau_{12}(t)\}}{\exp\{\tau_{11}(t)\} + \exp\{\tau_{12}(t)\} + 1}, \tag{II.3.89}$$

$$\Upsilon_{13}(t) = 1 - \sum_{j=1}^{K-1} \Upsilon_{1j}(t), \tag{II.3.90}$$

avec les transformations inverses

$$\tau_{11}(t) = \log\left(\frac{\Upsilon_{11}(t)}{\Upsilon_{13}(t)}\right) = \log\left(\frac{\Upsilon_{11}(t)}{1 - (\Upsilon_{11}(t) + \Upsilon_{12}(t))}\right), \tag{II.3.91}$$

$$\tau_{12}(t) = \log\left(\frac{\Upsilon_{12}(t)}{\Upsilon_{13}(t)}\right) = \log\left(\frac{\Upsilon_{12}(t)}{1 - (\Upsilon_{11}(t) + \Upsilon_{12}(t))}\right). \tag{II.3.92}$$

II.3.5.4. Matrice jacobienne de la reparamétrisation

Un enjeu qui survient lorsque l'on exprime la vraisemblance en fonction des paramètres de travail est que le gradient qui doit être donné à la routine d'optimisation et qui sera retourné par celle-ci est en fontion de ces paramètres de travail non-contraints. Nous aurons donc besoin de pouvoir aisément passer d'une paramétrisation à l'autre afin d'accélérer l'optimisation, d'obtenir les erreurs-types sur nos paramètres estimés et de calculer les quantités nécessaires au test d'efficacité. C'est à ce moment qu'entre en jeu un outil mathématique très important, le **jacobien**.

Définition II.3.5 (Jacobien)**.** *Soit* $g\colon \mathbb{R}^n \mapsto \boldsymbol{\Theta} \subseteq \mathbb{R}^D$*, une fonction dérivable en* $\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \left(\tilde{\theta}_1, \ldots, \tilde{\theta}_n\right) \in \mathbb{R}^n$ *:*

$$g : \begin{pmatrix} \tilde{\theta}_1 \\ \vdots \\ \tilde{\theta}_n \end{pmatrix} \mapsto \begin{pmatrix} g_1\left(\tilde{\theta}_1, \ldots, \tilde{\theta}_n\right) \\ \vdots \\ g_D\left(\tilde{\theta}_1, \ldots, \tilde{\theta}_n\right) \end{pmatrix}. \tag{II.3.93}$$

La ***matrice jacobienne*** *de g en $\tilde{\boldsymbol{\theta}}$, notée $\mathcal{J}_g\big(\tilde{\boldsymbol{\theta}}\big)$, est définie par*

$$\mathcal{J}_g\big(\tilde{\boldsymbol{\theta}}\big) \stackrel{\text{def}}{=} \frac{\partial g}{\partial \tilde{\boldsymbol{\theta}}'}\big(\tilde{\boldsymbol{\theta}}\big) = \begin{bmatrix} \frac{\partial g}{\partial \tilde{\theta}_1}\big(\tilde{\boldsymbol{\theta}}\big) & \dots & \frac{\partial g}{\partial \tilde{\theta}_n}\big(\tilde{\boldsymbol{\theta}}\big) \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial \tilde{\theta}_1}\big(\tilde{\boldsymbol{\theta}}\big) & \dots & \frac{\partial g_1}{\partial \tilde{\theta}_n}\big(\tilde{\boldsymbol{\theta}}\big) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D}{\partial \tilde{\theta}_1}\big(\tilde{\boldsymbol{\theta}}\big) & \dots & \frac{\partial g_D}{\partial \tilde{\theta}_n}\big(\tilde{\boldsymbol{\theta}}\big) \end{bmatrix} \in \mathbb{R}^{n \times D}. \quad \text{(II.3.94)}$$

L'élément au croisement de la i^{e} ligne et de la j^{e} colonne de la matrice $\mathcal{J}_g\big(\tilde{\boldsymbol{\theta}}\big)$ est noté $\big[\mathcal{J}_g\big(\tilde{\boldsymbol{\theta}}\big)\big]_{ij}$ et correspond à

$$\big[\mathcal{J}_g\big(\tilde{\boldsymbol{\theta}}\big)\big]_{ij} = \frac{\partial g_i}{\partial \tilde{\theta}_j}\big(\tilde{\boldsymbol{\theta}}\big). \quad \text{(II.3.95)}$$

La matrice $\mathcal{J}_g\big(\tilde{\boldsymbol{\theta}}\big)$ n'est fonction que de $\tilde{\boldsymbol{\theta}}$.

Remarque II.3.7. Le jacobien de la log-vraisemblance correspond à la transposée de son gradient :

$$\mathcal{J}_{\ell_T}(\boldsymbol{\theta}) = \boldsymbol{\nabla}' \ell_T(\boldsymbol{\theta}) = \mathcal{S}'(\boldsymbol{\theta} \mid x_{1:T}). \quad \text{(II.3.96)}$$

De plus, la matrice hessienne de la log-vraisemblance correspond au jacobien de son gradient :

$$\mathcal{H}_{\ell_T}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial^2 \ell_T}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'}\left(\frac{\partial \ell_T}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})\right) = \frac{\partial \boldsymbol{\nabla} \ell_T}{\partial \boldsymbol{\theta}'}(\boldsymbol{\theta}) = \mathcal{J}_{\boldsymbol{\nabla} \ell_T}(\boldsymbol{\theta}). \quad \text{(II.3.97)}$$

Une expression pour le score par rapport aux paramètres naturels peut à présent être obtenue par l'entremise du jacobien de la transformation de reparamétrisation g.

Proposition II.3.3. *Soit $\ell \colon \boldsymbol{\Theta} \mapsto \mathbb{R}$, la fonction de log-vraisemblance $\ell_T(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T}))$, où $\boldsymbol{\theta} \stackrel{\text{def}}{=} (\theta_1, \dots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$, avec D le nombre de paramètres libres du modèle et $\boldsymbol{\Theta}$ l'espace des paramètres.*

Soit $\tilde{\ell} \colon \mathbb{R}^n \mapsto \mathbb{R}$, la fonction de log-vraisemblance reparamétrisée $\tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big) \stackrel{\text{def}}{=} \ell_T\Big(g\big(\tilde{\boldsymbol{\theta}}\big)\Big)$, où $\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \big(\tilde{\theta}_1, \dots, \tilde{\theta}_n\big)' \in \mathbb{R}^n$, avec n le nombre de paramètres de travail et $g \colon \mathbb{R}^n \mapsto \boldsymbol{\Theta}$, une fonction surjective dérivable en $\tilde{\boldsymbol{\theta}}$.

Le gradient de $\tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big)$ est alors

$$\boldsymbol{\nabla} \tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big) \stackrel{\text{def}}{=} \frac{\partial \tilde{\ell}_T}{\partial \tilde{\boldsymbol{\theta}}}\big(\tilde{\boldsymbol{\theta}}\big) = \mathcal{J}'_g\big(\tilde{\boldsymbol{\theta}}\big) \cdot \boldsymbol{\nabla} \ell_T(\boldsymbol{\theta}), \quad \text{(II.3.98)}$$

avec $\mathcal{J}'_g\big(\tilde{\boldsymbol{\theta}}\big) \in \mathbb{R}^{D \times n}$.

Démonstration. Par la règle de la dérivation en chaîne, nous savons que

$$h(x) = f(g(x)) \implies \boldsymbol{\jmath}_h(x) = \boldsymbol{\jmath}_f(g(x)) \cdot \boldsymbol{\jmath}_g(x). \tag{II.3.99}$$

Ainsi, nous avons directement que

$$\tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big) = \ell_T\Big(g\big(\tilde{\boldsymbol{\theta}}\big)\Big) \implies \boldsymbol{\jmath}_{\tilde{\ell}_T}\big(\tilde{\boldsymbol{\theta}}\big) = \boldsymbol{\jmath}_{\ell_T}(\boldsymbol{\theta}) \cdot \boldsymbol{\jmath}_g\big(\tilde{\boldsymbol{\theta}}\big), \tag{II.3.100}$$

avec $\boldsymbol{\theta} = g\big(\tilde{\boldsymbol{\theta}}\big)$.

Puisque la transposée est une fonction bijective, que $(A \cdot B)' = B' \cdot A'$ et que $\boldsymbol{\jmath}_{\tilde{\ell}_T}\big(\tilde{\boldsymbol{\theta}}\big) = \boldsymbol{\nabla}'\tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big)$ et $\boldsymbol{\jmath}_g\big(\tilde{\boldsymbol{\theta}}\big) = \boldsymbol{\nabla}'\ell_T(\boldsymbol{\theta})$ (Remarque II.3.7), nous avons finalement que

$$\boldsymbol{\nabla}\tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big) = \boldsymbol{\jmath}'_g\big(\tilde{\boldsymbol{\theta}}\big) \cdot \boldsymbol{\nabla}\ell_T(\boldsymbol{\theta}). \tag{II.3.101}$$

□

Proposition II.3.4. *Soit $\ell\colon \boldsymbol{\Theta} \mapsto \mathbb{R}$, la fonction de log-vraisemblance $\ell_T(\boldsymbol{\theta}) \overset{\text{def}}{=} \log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T}))$, où $\boldsymbol{\theta} \overset{\text{def}}{=} (\theta_1, \ldots, \theta_D)' \in \boldsymbol{\Theta} \subseteq \mathbb{R}^D$, avec D le nombre de paramètres libres du modèle et $\boldsymbol{\Theta}$ l'espace des paramètres.*

Soit $\tilde{\ell}\colon \mathbb{R}^n \mapsto \mathbb{R}$, la fonction de log-vraisemblance reparamétrisée $\tilde{\ell}_T\big(\tilde{\boldsymbol{\theta}}\big) \overset{\text{def}}{=} \ell_T\Big(g\big(\tilde{\boldsymbol{\theta}}\big)\Big)$, où $\tilde{\boldsymbol{\theta}} \overset{\text{def}}{=} \big(\tilde{\theta}_1, \ldots, \tilde{\theta}_n\big)' \in \mathbb{R}^n$, avec n le nombre de paramètres de travail et $g\colon \mathbb{R}^n \mapsto \boldsymbol{\Theta}$, une fonction surjective dérivable en $\tilde{\boldsymbol{\theta}}$.

La matrice hessienne de $\tilde{\ell}_T\big(\check{\boldsymbol{\theta}}\big)$, avec $\widehat{\boldsymbol{\theta}} = g\big(\check{\boldsymbol{\theta}}\big)$ un point stationnaire de $\ell_T(\boldsymbol{\theta})$, est alors

$$\boldsymbol{\mathcal{H}}_{\tilde{\ell}_T}\big(\check{\boldsymbol{\theta}}\big) = \boldsymbol{\jmath}'_g\big(\check{\boldsymbol{\theta}}\big)\boldsymbol{\mathcal{H}}_{\ell_T}\big(\widehat{\boldsymbol{\theta}}\big)\boldsymbol{\jmath}_g\big(\check{\boldsymbol{\theta}}\big), \tag{II.3.102}$$

avec $\boldsymbol{\jmath}'_g\big(\check{\boldsymbol{\theta}}\big) \in \mathbb{R}^{D \times n}$.

Démonstration. Voir Augustyniak [2018](p. 38-39). □

En imposant pour le moment l'homogénéité à la matrice de transition, nous avons que la transformation g est de la forme

$$g : \begin{pmatrix} \tilde{\mu}_1 \\ \vdots \\ \tilde{\mu}_K \\ \eta_1 \\ \vdots \\ \eta_K \\ \upsilon_1 \\ \vdots \\ \upsilon_{K-1} \\ \tau_{11} \\ \vdots \\ \tau_{1(K-1)} \\ \vdots \\ \tau_{K1} \\ \vdots \\ \tau_{K(K-1)} \end{pmatrix} \mapsto \begin{pmatrix} g_1(\tilde{\mu}_1) \\ \vdots \\ g_K(\tilde{\mu}_K) \\ g_{K+1}(\eta_1) \\ \vdots \\ g_{2K}(\eta_K) \\ g_{2K+1}(\upsilon_1, \ldots, \upsilon_{K-1}) \\ \vdots \\ g_{3K-1}(\upsilon_1, \ldots, \upsilon_{K-1}) \\ g_{3K}\Big(\tau_{11}, \ldots, \tau_{1(K-1)}\Big) \\ \vdots \\ g_{4K-2}\Big(\tau_{11}, \ldots, \tau_{1(K-1)}\Big) \\ \vdots \\ g_{K^2+K}\Big(\tau_{K1}, \ldots, \tau_{K(K-1)}\Big) \\ \vdots \\ g_{K^2+2K-1}\Big(\tau_{K1}, \ldots, \tau_{K(K-1)}\Big) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \\ \sigma_1 \\ \vdots \\ \sigma_K \\ \mathfrak{u}_1(1) \\ \vdots \\ \mathfrak{u}_{K-1}(1) \\ \gimel_{11} \\ \vdots \\ \gimel_{1(K-1)} \\ \vdots \\ \gimel_{K1} \\ \vdots \\ \gimel_{K(K-1)} \end{pmatrix}. \tag{II.3.103}$$

Afin d'illustrer la démarche de construction de la matrice jacobienne, nous revenons à notre exemple II.3.1, où nous considérions un modèle de Markov à variables latentes homogène à $K = 3$ régimes.

Exemple II.3.3 (Suite de l'exemple II.3.1 : Construction du jacobien)**.** *Soit les transformations présentées aux équations* (II.3.80)*,* (II.3.82) *et* (II.3.84) *(p.169) qui combinées forme la transformation* $g\Big(\tilde{\boldsymbol{\theta}}\Big)$ *suivante pour une modèle homogène à* $K = 3$ *régimes :*

$$g : \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \tilde{\mu}_3 \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \upsilon_1 \\ \upsilon_2 \\ \tau_{11} \\ \tau_{12} \\ \tau_{21} \\ \tau_{22} \\ \tau_{31} \\ \tau_{32} \end{pmatrix} \mapsto \begin{pmatrix} g_1(\tilde{\mu}_1) \\ g_2(\tilde{\mu}_2) \\ g_3(\tilde{\mu}_3) \\ g_4(\eta_1) \\ g_5(\eta_2) \\ g_6(\eta_3) \\ g_7(\upsilon_1, \upsilon_2) \\ g_8(\upsilon_1, \upsilon_2) \\ g_9(\tau_{11}, \tau_{12}) \\ g_{10}(\tau_{11}, \tau_{12}) \\ g_{11}(\tau_{21}, \tau_{22}) \\ g_{12}(\tau_{21}, \tau_{32}) \\ g_{13}(\tau_{31}, \tau_{32}) \\ g_{14}(\tau_{31}, \tau_{32}) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \mathfrak{u}_1(1) \\ \mathfrak{u}_2(1) \\ \Upsilon_{11} \\ \Upsilon_{12} \\ \Upsilon_{21} \\ \Upsilon_{22} \\ \Upsilon_{31} \\ \Upsilon_{32} \end{pmatrix} \tag{II.3.104}$$

La matrice jacobienne est dans ce cas de la forme

$$\mathfrak{J}_g\big(\tilde{\boldsymbol{\theta}}\big) = \begin{bmatrix} A_{3\times 3} & & & \mathbf{0} \\ & B_{3\times 3} & & \\ & & C_{2\times 2} & \\ \mathbf{0} & & & D_{6\times 6} \end{bmatrix}, \tag{II.3.105}$$

avec

$$A = \begin{bmatrix} \frac{\partial g_1}{\partial \tilde{\mu}_1}(\tilde{\mu}_1) & \frac{\partial g_1}{\partial \tilde{\mu}_2}(\tilde{\mu}_1) & \frac{\partial g_1}{\partial \tilde{\mu}_3}(\tilde{\mu}_1) \\ \frac{\partial g_2}{\partial \tilde{\mu}_1}(\tilde{\mu}_2) & \frac{\partial g_2}{\partial \tilde{\mu}_2}(\tilde{\mu}_2) & \frac{\partial g_2}{\partial \tilde{\mu}_3}(\tilde{\mu}_2) \\ \frac{\partial g_3}{\partial \tilde{\mu}_1}(\tilde{\mu}_3) & \frac{\partial g_3}{\partial \tilde{\mu}_2}(\tilde{\mu}_3) & \frac{\partial g_3}{\partial \tilde{\mu}_3}(\tilde{\mu}_3) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \tilde{\mu}_1}\tilde{\mu}_1 & \frac{\partial}{\partial \tilde{\mu}_2}\tilde{\mu}_1 & \frac{\partial}{\partial \tilde{\mu}_3}\tilde{\mu}_1 \\ \frac{\partial}{\partial \tilde{\mu}_1}\tilde{\mu}_2 & \frac{\partial}{\partial \tilde{\mu}_2}\tilde{\mu}_2 & \frac{\partial}{\partial \tilde{\mu}_3}\tilde{\mu}_2 \\ \frac{\partial}{\partial \tilde{\mu}_1}\tilde{\mu}_3 & \frac{\partial}{\partial \tilde{\mu}_2}\tilde{\mu}_3 & \frac{\partial}{\partial \tilde{\mu}_3}\tilde{\mu}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{II.3.106}$$

$$\begin{aligned} B &= \begin{bmatrix} \frac{\partial g_4}{\partial \eta_1}(\eta_1) & \frac{\partial g_4}{\partial \eta_2}(\eta_1) & \frac{\partial g_4}{\partial \eta_3}(\eta_1) \\ \frac{\partial g_5}{\partial \eta_1}(\eta_2) & \frac{\partial g_5}{\partial \eta_2}(\eta_2) & \frac{\partial g_5}{\partial \eta_3}(\eta_2) \\ \frac{\partial g_6}{\partial \eta_1}(\eta_3) & \frac{\partial g_6}{\partial \eta_2}(\eta_3) & \frac{\partial g_6}{\partial \eta_3}(\eta_3) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \eta_1}[\exp\{\eta_1\}] & \frac{\partial}{\partial \eta_2}[\exp\{\eta_1\}] & \frac{\partial}{\partial \eta_3}[\exp\{\eta_1\}] \\ \frac{\partial}{\partial \eta_1}[\exp\{\eta_2\}] & \frac{\partial}{\partial \eta_2}[\exp\{\eta_2\}] & \frac{\partial}{\partial \eta_3}[\exp\{\eta_2\}] \\ \frac{\partial}{\partial \eta_1}[\exp\{\eta_3\}] & \frac{\partial}{\partial \eta_2}[\exp\{\eta_3\}] & \frac{\partial}{\partial \eta_3}[\exp\{\eta_3\}] \end{bmatrix} \\ &= \begin{bmatrix} \exp\{\eta_1\} & 0 & 0 \\ 0 & \exp\{\eta_2\} & 0 \\ 0 & 0 & \exp\{\eta_3\} \end{bmatrix}, \end{aligned} \tag{II.3.107}$$

$$C = \begin{bmatrix} \frac{\partial g_7}{\partial \upsilon_1}(\upsilon_1, \upsilon_2) & \frac{\partial g_7}{\partial \upsilon_2}(\upsilon_1, \upsilon_2) \\ \frac{\partial g_8}{\partial \upsilon_1}(\upsilon_1, \upsilon_2) & \frac{\partial g_8}{\partial \upsilon_2}(\upsilon_1, \upsilon_2) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \upsilon_1}\left[\frac{\exp\{\upsilon_1\}}{1+\exp\{\upsilon_1\}+\exp\{\upsilon_2\}}\right] & \frac{\partial}{\partial \upsilon_2}\left[\frac{\exp\{\upsilon_1\}}{1+\exp\{\upsilon_1\}+\exp\{\upsilon_2\}}\right] \\ \frac{\partial}{\partial \upsilon_1}\left[\frac{\exp\{\upsilon_2\}}{1+\exp\{\upsilon_1\}+\exp\{\upsilon_2\}}\right] & \frac{\partial}{\partial \upsilon_2}\left[\frac{\exp\{\upsilon_2\}}{1+\exp\{\upsilon_1\}+\exp\{\upsilon_2\}}\right] \end{bmatrix}$$

$$= \begin{bmatrix} \mathfrak{u}_1(1)\Big[\frac{1+\exp\{v_2\}}{1+\exp\{v_1\}+\exp\{v_2\}}\Big] & -\frac{\exp\{v_1\}\exp\{v_2\}}{(1+\exp\{v_1\}+\exp\{v_2\})^2} \\ -\frac{\exp\{v_1\}\exp\{v_2\}}{(1+\exp\{v_1\}+\exp\{v_2\})^2} & \mathfrak{u}_2(1)\Big[\frac{1+\exp\{v_1\}}{1+\exp\{v_1\}+\exp\{v_2\}}\Big] \end{bmatrix}, \tag{II.3.108}$$

$$D = \begin{bmatrix} \frac{\partial g_9}{\partial \tau_{11}}(\tau_{11},\tau_{12}) & \frac{\partial g_9}{\partial \tau_{12}}(\tau_{11},\tau_{12}) & 0 & 0 & 0 & 0 \\ \frac{\partial g_{10}}{\partial \tau_{11}}(\tau_{11},\tau_{12}) & \frac{\partial g_{10}}{\partial \tau_{12}}(\tau_{11},\tau_{12}) & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial g_{11}}{\partial \tau_{21}}(\tau_{21},\tau_{22}) & \frac{\partial g_{11}}{\partial \tau_{22}}(\tau_{21},\tau_{22}) & 0 & 0 \\ 0 & 0 & \frac{\partial g_{12}}{\partial \tau_{21}}(\tau_{21},\tau_{22}) & \frac{\partial g_{12}}{\partial \tau_{22}}(\tau_{21},\tau_{22}) & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\partial g_{13}}{\partial \tau_{31}}(\tau_{31},\tau_{32}) & \frac{\partial g_{13}}{\partial \tau_{32}}(\tau_{31},\tau_{32}) \\ 0 & 0 & 0 & 0 & \frac{\partial g_{14}}{\partial \tau_{31}}(\tau_{31},\tau_{32}) & \frac{\partial g_{14}}{\partial \tau_{32}}(\tau_{31},\tau_{32}) \end{bmatrix}$$

$$= \begin{bmatrix} D_1 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 \\ 0 & 0 & D_2 & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & 0 & D_3 \\ 0 & 0 & 0 & 0 & \end{bmatrix},$$

$$D_i = \begin{bmatrix} \frac{\partial g_{7+2i}}{\partial \tau_{i1}}(\tau_{i1},\tau_{i2}) & \frac{\partial g_{7+2i}}{\partial \tau_{i2}}(\tau_{i1},\tau_{i2}) \\ \frac{\partial g_{8+2i}}{\partial \tau_{i1}}(\tau_{i1},\tau_{i2}) & \frac{\partial g_{8+2i}}{\partial \tau_{i2}}(\tau_{i1},\tau_{i2}) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \tau_{i1}}\Big[\frac{\exp\{\tau_{i1}\}}{1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\}}\Big] & \frac{\partial}{\partial \tau_{i2}}\Big[\frac{\exp\{\tau_{i1}\}}{1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\}}\Big] \\ \frac{\partial}{\partial \tau_{i1}}\Big[\frac{\exp\{\tau_{i2}\}}{1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\}}\Big] & \frac{\partial}{\partial \tau_{i2}}\Big[\frac{\exp\{\tau_{i2}\}}{1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\}}\Big] \end{bmatrix}$$

$$= \begin{bmatrix} g_{7+2i}(\tau_{i1},\tau_{i2})\Big[\frac{1+\exp\{\tau_{i2}\}}{1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\}}\Big] & \frac{-\exp\{\tau_{i1}\}\exp\{\tau_{i2}\}}{(1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\})^2} \\ \frac{-\exp\{\tau_{i1}\}\exp\{\tau_{i2}\}}{(1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\})^2} & g_{8+2i}(\tau_{i1},\tau_{i2})\Big[\frac{1+\exp\{\tau_{i1}\}}{1+\exp\{\tau_{i1}\}+\exp\{\tau_{i2}\}}\Big] \end{bmatrix}, \quad i \in \{1,2,3\}. \tag{II.3.109}$$

Il est bien connu qu'une fonction peut avoir plus d'un sommet, un exemple évident de cette caractéristique étant la fonction trigonométrique $\sin(x)$ qui est périodique et admet une infinité d'extrema sur les $\mathbb{R}$. Il se peut donc que le point trouvé ne soit qu'un maximum local et que le vecteur de paramètres obtenu ne soit pas l'EMV, même si le gradient en ce point est nul. Cette considération sera discutée à la fin de la présente section, en plus du défi de choisir de bonnes valeurs initiales pour les paramètres dans la routine d'optimisation.

II.3.6. Efficacité, convergence et propriétés asymptotiques de l'EMV

Définition II.3.6 (Matrice d'information de Fisher)**.** *La* ***matrice d'information de Fisher****, notée* $\mathcal{F}_T(\boldsymbol{\theta})$*, est*

$$\mathcal{F}_T(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}\Big[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})\mathcal{S}'(\boldsymbol{\theta} \mid x_{1:T})\Big], \tag{II.3.110}$$

$$= \int \mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})\mathcal{S}'(\boldsymbol{\theta} \mid x_{1:T})\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})d_{x_{1:T}}, \tag{II.3.111}$$

où le score $\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})$ *correspond au* ***gradient*** *de log-vraisemblance en* $\boldsymbol{\theta}$ *(Définition II.3.2).*

La matrice $\mathcal{F}_T(\boldsymbol{\theta})$ *n'est fonction que de* $\boldsymbol{\theta}$*, l'échantillon étant considéré aléatoire.*

Proposition II.3.5. *Sous certaines conditions de différentiabilité sous le signe de l'intégrale, on obtient que*

$$\mathcal{F}_T(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{I}(\boldsymbol{\theta} \mid x_{1:T})] = \mathbb{E}\left[-\frac{\partial^2 \ell_T}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta})\right]. \tag{II.3.112}$$

Démonstration. Voir Ferguson [1996](p. 120). □

Remarque II.3.8. Puisque

$$\mathbb{V}ar[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})] = \mathbb{E}\Big[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})\mathcal{S}'(\boldsymbol{\theta} \mid x_{1:T})\Big] - \mathbb{E}[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})]\mathbb{E}[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})]'$$

et que

$$\begin{aligned}
\mathbb{E}[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})] &= \int \frac{\partial}{\partial\boldsymbol{\theta}}\log(\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})) \times \mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})d_{x_{1:T}}, \\
&= \int \frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})d_{x_{1:T}}, \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\int \mathcal{P}_{\boldsymbol{\theta}}(x_{1:T})d_{x_{1:T}}, \\
&= \mathbf{0},
\end{aligned}$$

la matrice d'information de Fisher peut aussi être écrite comme la variance du score, soit

$$\mathcal{F}_T(\boldsymbol{\theta}) = \mathbb{V}ar[\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T})]. \tag{II.3.113}$$

Théorème II.3.2. *Soit* $x_{1:T}$*, un échantillon* iid *provenant de la densité* $\mathbb{P}[X_t \mid \boldsymbol{\theta_0}]$*, avec* $\boldsymbol{\theta_0}$ *le vecteur de paramètres réels ayant engendrés les observations. Sous certaines conditions dites de régularité, la théorie asymptotique du maximum de vraisemblance stipule que dans tout voisinage suffisamment petit de* $\boldsymbol{\theta_0}$*, lorsque* T *est suffisamment grand, il existe une solution unique à l'équation* $\mathcal{S}(\boldsymbol{\theta} \mid x_{1:T}) = 0$*, notée* $\widehat{\boldsymbol{\theta}}$*, qui maximise localement la log-vraisemblance.*

De plus, cette solution, que l'on appelle l'EMV, est fortement convergente

$$\widehat{\boldsymbol{\theta}} \overset{p.s.}{\to} \boldsymbol{\theta_0}, \tag{II.3.114}$$

et aymptotiquement efficace

$$\sqrt{T}\Big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\Big) \overset{d}{\to} \mathcal{N}\Big(0, [\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0})]^{-1}\Big). \tag{II.3.115}$$

Par conséquent, la distribution échantillonale de l'estimateur $\widehat{\boldsymbol{\theta}}$ peut être approximée par

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}\Big(\boldsymbol{\theta_0}, [\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0})]^{-1}\Big), \tag{II.3.116}$$

où l'inverse de la matrice d'information de Fisher en $\boldsymbol{\theta_0}$, soit $[\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0})]^{-1}$, est appelée la ***matrice de variance-covariance asymptotique*** *de $\widehat{\boldsymbol{\theta}}$. Il est donc possible d'approximer les erreurs-types sur les éléments de l'estimateur $\widehat{\boldsymbol{\theta}}$ par*

$$\sqrt{\operatorname{diag}\Big([\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0})]^{-1}\Big)}. \tag{II.3.117}$$

Démonstration. Nous recommandons les références proposées dans Augustyniak [2018](p. 49), soit le chapitre 18 de Ferguson [1996] pour le cas général et Redner et Walker [1984] pour les particularités liées aux modèles de mélange. □

Remarque II.3.9. Au chapitre 7 de Ferguson [1996], il est aussi démontré que si la transformation g est différentiable en $\boldsymbol{\theta}$ et que $\widehat{\boldsymbol{\theta}}$ est un estimateur asymptotiquement efficace de $\boldsymbol{\theta}$ tel que

$$\sqrt{T}\Big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big) \overset{d}{\to} \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{II.3.118}$$

alors

$$\sqrt{T}\Big(g\Big(\widehat{\boldsymbol{\theta}}\Big) - g(\boldsymbol{\theta})\Big) \overset{d}{\to} \mathcal{N}\Big(0, \boldsymbol{\mathcal{J}}_g(\boldsymbol{\theta})\, \boldsymbol{\Sigma}\, \boldsymbol{\mathcal{J}}_g'(\boldsymbol{\theta})\Big). \tag{II.3.119}$$

En pratique cependant, il n'est jamais possible de connaître les paramètres réels $\boldsymbol{\theta_0}$ de notre modèle. Il n'est de surcroît que très rare de pouvoir obtenir analytiquement la matrice d'infomation de Fisher $\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0})$ pour la même raison. Puisque nous avons que $\widehat{\boldsymbol{\theta}} \overset{p.s.}{\to} \boldsymbol{\theta_0}$, il est commun dans la pratique d'utiliser l'approximation

$$\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0}) \approx \boldsymbol{\mathcal{I}}\Big(\widehat{\boldsymbol{\theta}} \;\Big|\; x_{1:T}\Big) \tag{II.3.120}$$

afin d'obtenir une estimation de l'erreur-type sur nos paramètres estimées. Similairement, il est aussi possible d'utiliser l'approximation

$$\boldsymbol{\mathcal{F}}_T(\boldsymbol{\theta_0}) \approx \sum_{t=1}^{T} \boldsymbol{\mathcal{S}}\Big(\widehat{\boldsymbol{\theta}} \;\Big|\; x_t\Big) \boldsymbol{\mathcal{S}}'\Big(\widehat{\boldsymbol{\theta}} \;\Big|\; x_t\Big). \tag{II.3.121}$$

II.3.7. Influence de l'initialisation des paramètres

La dernière préoccupation d'importance est que la fonction à maximiser admet potentiellement plus d'un maximum local sur $\boldsymbol{\Theta}$. La fonction de vraisemblance d'un *HMM* est sensible à ce genre de problème et le résultat de l'optimisation est donc potentiellement affecté par le choix des paramètres initiaux donnés à la routine d'optimisation. Une manière de se protéger contre cette situation est d'estimer le modèle avec plusieurs combinaisons de valeurs initiales et de voir si le maximum obtenu est le même. Si les valeurs obtenues ne sont pas toutes identiques, il faudra choisir une méthode afin de départager les vecteurs de paramètres obtenus et selectionner la combinaison optimale selon des critères préétablis.

Appendice III

Démonstrations

III.1. Distribution marginale d'une chaîne de Markov 2.5

Démonstration. Sans perte de généralité, prenons la probabilité $\mathfrak{u}_k(t+n) = \mathbb{P}[C_{t+n} = k]$:

$$\begin{aligned}
\mathfrak{u}_k(t+n) &= \sum_{i=1}^{K} \mathbb{P}[C_{t+n} = k, C_t = i] && \text{Probabilités totales,} \\
&= \sum_{i=1}^{K} \mathbb{P}[C_t = i]\mathbb{P}[C_{t+n} = k \mid C_t = i] && \text{Bayes,} \\
&= \sum_{i=1}^{K} \mathfrak{u}_i(t) \times \gimel_{ik}^{(n)}(t), \\
&= \vec{\mathfrak{u}}(t) \times \text{la } k^{\text{e}} \text{ colonne de } \mathbf{\Gamma}_t^{(n)}.
\end{aligned}$$

□

III.2. Fonction d'autocovariance d'une chaîne de Markov I.6.1

Démonstration. Tout d'abord, nous savons par définition de la covariance que

$$\mathbb{C}ov[C_t, C_{t+h}] = \mathbb{E}[C_t C_{t+h}] - \mathbb{E}[C_t]\mathbb{E}[C_{t+h}].$$

Maintenant, nous avons que

$$\begin{aligned}\mathbb{E}[C_tC_{t+h}] &= \mathbb{E}[C_t\ \mathbb{E}[C_{t+h} \mid C_t]],\\ &= \sum_{i=1}^{K}(\upsilon_i \times \mathbb{E}[C_{t+h} \mid C_t = \upsilon_i])\mathbb{P}[C_t = \upsilon_i],\\ &= \sum_{i=1}^{K}\upsilon_i\left(\sum_{j=1}^{K}\upsilon_j \times \Upsilon_{ij}^{(h)}(t)\right)\mathfrak{u}_i(t),\\ &= \sum_{i=1}^{K}(\mathfrak{u}_i(t)\upsilon_i)\left(\sum_{j=1}^{K}\Upsilon_{ij}^{(h)}(t)\upsilon_j\right).\end{aligned}$$

De plus, il est possible d'écrire cette équation en notation matricielle en utilisant le fait que

$$\vec{\mathfrak{u}}(t)\Upsilon = (\mathfrak{u}_1(t)\upsilon_1\ \ldots\ \mathfrak{u}_K(t)\upsilon_K)$$

et que

$$\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\upsilon} = \left(\sum_{j=1}^{K}\Upsilon_{1j}^{(h)}(t)\upsilon_j\ \ldots\ \sum_{j=1}^{K}\Upsilon_{Kj}^{(h)}(t)\upsilon_j\right)',$$

pour obtenir que

$$\mathbb{E}[C_tC_{t+h}] = \vec{\mathfrak{u}}(t)\boldsymbol{\Upsilon}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\upsilon}.$$

En utilisant maintenant le résultat en (I.6.2), nous obtenons finalement le résultat en (I.6.4)

$$\begin{aligned}\mathbb{C}ov[C_t, C_{t+h}] &= \mathbb{E}[C_tC_{t+h}] - \mathbb{E}[C_t]\mathbb{E}[C_{t+h}],\\ &= \vec{\mathfrak{u}}(t)\boldsymbol{\Upsilon}\boldsymbol{\Gamma}_t^{(h)}\boldsymbol{\upsilon} - (\vec{\mathfrak{u}}(t)\boldsymbol{\upsilon}) \times (\vec{\mathfrak{u}}(t+h)\boldsymbol{\upsilon}).\end{aligned}$$

Le résultat (I.6.5) est obtenu directement en remplaçant les distributions marginales par la distribution stationnaire et la matrice de transition par son homologue homogène

$$\begin{aligned}\mathbb{C}ov[C_t, C_{t+h}] &= \vec{\pi}\boldsymbol{\Upsilon}\boldsymbol{\Gamma}^h\boldsymbol{\upsilon} - (\vec{\pi}\boldsymbol{\upsilon}) \times (\vec{\pi}\boldsymbol{\upsilon}),\\ &= \vec{\pi}\boldsymbol{\Upsilon}\boldsymbol{\Gamma}^h\boldsymbol{\upsilon} - (\vec{\pi}\boldsymbol{\upsilon})^2.\end{aligned}$$

Les résultats concernant la variance de C_t sont obtenus directement, puisque $\boldsymbol{\Gamma}_t^{(0)} = \mathbf{I}_K, \quad \forall\, t \in \mathbb{N}$. □

III.3. Fonction d'autocorrélation d'une chaîne de Markov ergodique I.6.2

DÉMONSTRATION. Tout d'abord, puisque la chaîne est stationnaire, l'autocorrélation ne dépend que du délais h. Ainsi, elle est définie comme suit :

$$\mathbb{C}\text{orr}[C_t, C_{t+h}] = \frac{\mathbb{C}ov[C_t, C_{t+h}]}{\mathbb{C}ov[C_t, C_t]}. \tag{III.3.1}$$

Par la proposition I.6.1, nous savons que la fonction d'autocovariance d'une chaîne de Markov ergodique est

$$\begin{aligned}\mathbb{C}ov[C_t, C_{t+h}] &= \vec{\pi}\Upsilon\Gamma^h\boldsymbol{v} - (\vec{\pi}\boldsymbol{v})^2, \\ &= \vec{\pi}\Upsilon Q\Lambda^h Q^{-1}\boldsymbol{v} - (\vec{\pi}\boldsymbol{v})^2 \end{aligned}\tag{III.3.2}$$

Nous savons de plus que puisque la première valeur propre de Γ est $\lambda_1 = 1$, la première colonne de Q est $\mathbf{1}_K$ et la première ligne de Q^{-1} correspond au vecteur $\vec{\pi}$. Ainsi, la première colonne de $Q\Lambda^h$ correspond au vecteur unité $\mathbf{1}_K$ et puisque $\Upsilon\mathbf{1}_K = \boldsymbol{v}$, nous avons que le produit ligne par ligne autour de la valeur propre λ_1 correspond à

$$\vec{\pi}\Upsilon\mathbf{1}_K\vec{\pi}\boldsymbol{v} = (\vec{\pi}\boldsymbol{v})^2,$$

ce qui s'annule avec le second terme de (III.3.2). Il ne reste donc que les lignes de 2 à K du produit matricielle, ce qui peut être réécrit sous la forme

$$\mathbb{C}ov[C_t, C_{t+h}] = \sum_{i=2}^{K} a_i b_i \lambda_i^h. \tag{III.3.3}$$

En réécrivant la fonction d'autocorrélation avec ce résultat, nous obtenons

$$\mathbb{C}\text{orr}[C_t, C_{t+h}] = \frac{\mathbb{C}ov[C_t, C_{t+h}]}{\mathbb{C}ov[C_t, C_{t+h}]|_{h=0}}, \tag{III.3.4}$$

$$= \frac{\sum_{i=2}^{K} a_i b_i \lambda_i^h}{\sum_{i=2}^{K} a_i b_i}. \tag{III.3.5}$$

□

III.4. Variance d'une transformation de la série observée dans un *HMM*, équation matricielle (I.7.3) (p.138)

DÉMONSTRATION. Sans perte de généralité, prenons $K = 3$,

$$\sum_{i=2}^{K}\sum_{j=1}^{i-1}\mu_{g|i}\ \mathfrak{u}_i(t)\,\mu_{g|j}\ \mathfrak{u}_j(t)\Big|_{K=3} = \underbrace{\mu_{g|2}\ \mathfrak{u}_2(t)}_{\textcircled{2}}\underbrace{\mu_{g|1}\ \mathfrak{u}_1(t)}_{\textcircled{1}} + \underbrace{\mu_{g|3}\ \mathfrak{u}_3(t)}_{\textcircled{3}}\Big(\mu_{g|1}\ \mathfrak{u}_1(t) + \mu_{g|2}\ \mathfrak{u}_2(t)\Big),$$

$$= \textcircled{1}\Big(\textcircled{2} + \textcircled{3}\Big) + \textcircled{2}\textcircled{3},$$

$$= \underbrace{\Big[\textcircled{1}\Big(\textcircled{1} + \textcircled{2} + \textcircled{3}\Big) + \textcircled{2}\Big(\textcircled{1} + \textcircled{2} + \textcircled{3}\Big) + \textcircled{3}\Big(\textcircled{1} + \textcircled{2} + \textcircled{3}\Big)\Big]}_{\star}$$

$$- \underbrace{\Big[\textcircled{1}\textcircled{1} + \textcircled{2}\textcircled{1} + \textcircled{2}\textcircled{2} + \textcircled{3}\textcircled{1} + \textcircled{3}\textcircled{2} + \textcircled{3}\textcircled{3}\Big]}_{\star\star}.$$

Afin d'alléger la notation des résultats intermédiaires, nous noterons

$$\vec{\Omega}_g^{(t)} \stackrel{\text{def}}{=} \vec{\mathfrak{u}}(t)\odot\boldsymbol{\mu}_g{}' = \begin{bmatrix}\textcircled{1} & \textcircled{2} & \textcircled{3}\end{bmatrix}. \tag{III.4.1}$$

D'abord, la première partie de cette équation peut être réécrite sous forme matricielle comme suit

$$\star = \begin{bmatrix}\textcircled{1} & \textcircled{2} & \textcircled{3}\end{bmatrix}\begin{bmatrix}\textcircled{1} & \textcircled{2} & \textcircled{3}\\ \textcircled{1} & \textcircled{2} & \textcircled{3}\\ \textcircled{1} & \textcircled{2} & \textcircled{3}\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}, \tag{III.4.2}$$

$$= \begin{bmatrix}\textcircled{1} & \textcircled{2} & \textcircled{3}\end{bmatrix}\left(\begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}\textcircled{1} & \textcircled{2} & \textcircled{3}\end{bmatrix}\right)\begin{bmatrix}1\\1\\1\end{bmatrix}, \tag{III.4.3}$$

$$= \vec{\Omega}_g^{(t)}\Big(\mathbf{1}_3\vec{\Omega}_g^{(t)}\Big)\mathbf{1}_3. \tag{III.4.4}$$

Maintenant, nous remarquons que les termes qui doivent être retirés à cette équation correspondent à ceux en dessous et incluant la diagonale de la matrice

$$\begin{bmatrix}\textcircled{1} & \textcircled{2} & \textcircled{3}\\ \textcircled{1} & \textcircled{2} & \textcircled{3}\\ \textcircled{1} & \textcircled{2} & \textcircled{3}\end{bmatrix}, \tag{III.4.5}$$

avec ces termes superflus en rouge. Nous aurons donc usage de la matrice diagonale inférieure

$$\llcorner_3 \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

afin de retirer ces termes. Ainsi,

$$\star\star = \begin{bmatrix} \textcircled{1} & \textcircled{2} & \textcircled{3} \end{bmatrix} \left(\begin{bmatrix} \textcircled{1} & \textcircled{2} & \textcircled{3} \\ \textcircled{1} & \textcircled{2} & \textcircled{3} \\ \textcircled{1} & \textcircled{2} & \textcircled{3} \end{bmatrix} \odot \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{III.4.6}$$

$$= \begin{bmatrix} \textcircled{1} & \textcircled{2} & \textcircled{3} \end{bmatrix} \begin{bmatrix} \textcircled{1} & 0 & 0 \\ \textcircled{1} & \textcircled{2} & 0 \\ \textcircled{1} & \textcircled{2} & \textcircled{3} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{III.4.7}$$

$$= \begin{bmatrix} \Big(\textcircled{1}\textcircled{1} + \textcircled{2}\textcircled{1} + \textcircled{3}\textcircled{1} \Big) & \Big(\textcircled{2}\textcircled{2} + \textcircled{3}\textcircled{2} \Big) & \textcircled{3}\textcircled{3} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{III.4.8}$$

et en notation matricielle,

$$\star\star = \vec{\Omega}_g^{(t)} \Big(\Big(\mathbf{1}_3 \vec{\Omega}_g^{(t)} \Big) \odot \llcorner_3 \Big) \mathbf{1}_3. \tag{III.4.9}$$

Et pour terminer, l'équation de départ avec $K = 3$ se réécrit sous la forme

$$\begin{aligned} \sum_{i=2}^{K} \sum_{j=1}^{i-1} \mu_{g|i}\ \mathfrak{u}_i(t)\, \mu_{g|j}\ \mathfrak{u}_j(t) \Big|_{K=3} &= \vec{\Omega}_g^{(t)} \Big(\mathbf{1}_3 \vec{\Omega}_g^{(t)} \Big) \mathbf{1}_3 - \vec{\Omega}_g^{(t)} \Big[\Big(\mathbf{1}_3 \vec{\Omega}_g^{(t)} \Big) \odot \llcorner_3 \Big] \mathbf{1}_3, \\ &= \vec{\Omega}_g^{(t)} \Big[\mathbf{1}_3 \vec{\Omega}_g^{(t)} - \Big(\mathbf{1}_3 \vec{\Omega}_g^{(t)} \Big) \odot \llcorner_3 \Big] \mathbf{1}_3, \\ &= \vec{\Omega}_g^{(t)} \Big[\Big(\mathbf{1}_3 \vec{\Omega}_g^{(t)} \Big) \odot \Big(\mathbf{1}_3 \mathbf{1}_3{}' - \llcorner_3 \Big) \Big] \mathbf{1}_3 \end{aligned} \tag{III.4.10}$$

avec le résultat général

$$\sum_{i=2}^{K} \sum_{j=1}^{i-1} \mu_{g|i} \mathfrak{u}_i(t)\, \mu_{g|j}\ \mathfrak{u}_j(t) = \vec{\Omega}_g^{(t)} \Big[\Big(\mathbf{1}_K \vec{\Omega}_g^{(t)} \Big) \odot \Big(\mathbf{1}_K \mathbf{1}_K{}' - \llcorner_K \Big) \Big] \mathbf{1}_K. \tag{III.4.11}$$

□

III.5. Démonstration du filtre de Kim

Nous voulons démontrer que, sous la spécification généralisée (p.28) d'un modèle de Markov à variables latentes, il est possible d'affirmer que

$$\mathbb{P}\Big[C_t = i \;\Big|\; C_{t+1} = j, x_{1:T}\Big] = \mathbb{P}\Big[C_t = i \;\Big|\; C_{t+1} = j, x_{1:t}\Big] \tag{III.5.1}$$

Démonstration. Tout d'abord, rappelons les hypothèses du *HMM* plus général présenté dans le cadre du filtre d'Hamilton à la section 3.6.1.1 :

$$\begin{cases} C_1 \sim p(c_1), \\ C_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t-1} = c_{1:t-1}) \sim p(c_t \mid c_{t-1}, x_{1:t-1}), \\ X_t \mid (X_{1:t-1} = x_{1:t-1}, C_{1:t} = c_{1:t}) \sim p(x_t \mid x_{1:t-1}, c_t). \end{cases} \tag{III.5.2}$$

En utilisant le théorème de Bayes, nous avons que

$$\begin{aligned} \mathbb{P}[c_t \mid c_{t+1}, x_{1:T}] &= \frac{\mathbb{P}[c_t, c_{t+1} \mid x_{1:T}]}{\mathbb{P}[c_{t+1} \mid x_{1:T}]}, \\ &= \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:T}]\mathbb{P}[c_t \mid x_{1:T}]}{\mathbb{P}[c_{t+1} \mid x_{1:T}]}, \\ &= \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:t}]\mathbb{P}[c_t \mid x_{1:T}]}{\mathbb{P}[c_{t+1} \mid x_{1:T}]}, \end{aligned} \tag{III.5.3}$$

Maintenant, en rappelant que si l'événement A est indépendant de l'événement B, alors $\mathbb{P}[A] = \mathbb{P}[A \mid B]$, nous développons $\mathbb{P}[c_{t+1} \mid x_{1:T}]$:

$$\begin{aligned} \mathbb{P}[c_{t+1} \mid x_{1:T}] &= \sum_{c_t} \mathbb{P}[c_{t+1}, c_t \mid x_{1:T}], \\ &= \sum_{c_t} \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:T}]}{\mathbb{P}[c_t \mid x_{1:T}]}, \\ &= \sum_{c_t} \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:t}]}{\mathbb{P}[c_t \mid x_{1:T}]}. \end{aligned}$$

En développant maintenant le dénominateur,

$$\begin{aligned}\mathbb{P}[c_t \mid x_{1:T}] &= \sum_{c_{t-1}} \mathbb{P}[c_t, c_{t-1} \mid x_{1:T}],\\ &= \sum_{c_{t-1}} \frac{\mathbb{P}[c_t \mid c_{t-1}, x_{1:T}]}{\mathbb{P}[c_{t-1} \mid x_{1:T}]},\\ &= \sum_{c_{t-1}} \frac{\mathbb{P}[c_t \mid c_{t-1}, x_{1:t-1}]}{\mathbb{P}[c_{t-1} \mid x_{1:T}]},\\ &= \sum_{c_{t-1}} \frac{\mathbb{P}[c_t \mid c_{t-1}, x_{1:t}]}{\mathbb{P}[c_{t-1} \mid x_{1:T}]},\end{aligned}$$

car puisque $c_t \mid c_{t-1}, x_{1:t-1}$ est indépendant de x_t, $\mathbb{P}[c_t \mid c_{t-1}, x_{1:t-1}] = \mathbb{P}[c_t \mid c_{t-1}, x_{1:t}]$ et ainsi de suite jusqu'à

$$\begin{aligned}\mathbb{P}[c_1 \mid x_{1:T}] &= \mathbb{P}[c_1] \quad \text{par hypothèse},\\ &= \mathbb{P}[c_1 \mid x_{1:t}].\end{aligned}$$

En remplaçant les valeurs obtenues récursivement, nous obtenons qu'à chaque point $k \leq t$,

$$\mathbb{P}[c_k \mid x_{1:T}] = \mathbb{P}[c_k \mid x_{1:t}]$$

et donc que

$$\begin{aligned}\mathbb{P}[c_{t+1} \mid x_{1:T}] &= \sum_{c_t} \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:t}]}{\mathbb{P}[c_t \mid x_{1:T}]},\\ &= \sum_{c_t} \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:t}]}{\mathbb{P}[c_t \mid x_{1:t}]},\\ &= \sum_{c_t} \mathbb{P}[c_{t+1}, c_t \mid x_{1:t}],\\ &= \mathbb{P}[c_{t+1} \mid x_{1:t}].\end{aligned}$$

En remplaçant ces quantités dans (III.5.3), nous obtenons donc

$$\begin{aligned}\mathbb{P}[C_t \mid c_{t+1}, x_{1:T}] &= \frac{\mathbb{P}[c_{t+1} \mid C_t, x_{1:t}]\mathbb{P}[c_t \mid x_{1:T}]}{\mathbb{P}[c_{t+1} \mid x_{1:T}]},\\ &= \frac{\mathbb{P}[c_{t+1} \mid c_t, x_{1:t}]\mathbb{P}[c_t \mid x_{1:t}]}{\mathbb{P}[c_{t+1} \mid x_{1:t}]},\\ &= \mathbb{P}[c_t \mid c_{t+1}, x_{1:t}].\end{aligned}$$

□

III.6. Démonstrations pour l'algorithme de Viterbi

III.6.1. Démontration de relation recursive de la proposition II.2.1

Nous allons prouver par induction que

$$\xi_{t+1}(j) = \mathbb{P}[x_{t+1} \mid x_{1:t}, C_{t+1} = j] \times \max_{i=1,\dots,K} \{\mathbb{P}[C_{t+1} = j \mid C_t = i, x_{1:T}]\xi_t(i)\}, \quad t = 1,\dots,T-1. \tag{III.6.1}$$

Démonstration. En $t = 1$, nous avons que

$$\begin{aligned}
\xi_2(j) &\stackrel{\text{def}}{=} \max_{c_1} \mathbb{P}[x_1, x_2, c_1, C_2 = j], \\
&= \max_{i=1,\dots,K} \mathbb{P}[x_1, x_2, C_1 = i, C_2 = j], \\
&= \max_{i=1,\dots,K} \{\mathbb{P}[x_1, C_1 = i]\mathbb{P}[x_2, C_2 = j \mid x_1, C_1 = i]\}, \\
&= \max_{i=1,\dots,K} \left\{ \underbrace{\mathbb{P}[x_1, C_1 = i]}_{\xi_1(i)} \mathbb{P}[C_2 = j \mid C_1 = i, x_1] \underbrace{\mathbb{P}[x_2 \mid x_1 C_2 = j]}_{\text{Indépendant de } i} \right\}, \\
&= \mathbb{P}[x_2 \mid x_1 C_2 = j] \max_{i=1,\dots,K} \{\mathbb{P}[C_2 = j \mid C_1 = i, x_1] \times \xi_1(i)\}.
\end{aligned}$$

Si nous supposons maintenant que

$$\xi_t(j) = \mathbb{P}[x_t \mid x_{1:t-1} C_t = j] \times \max_{i=1,\dots,K} \{\mathbb{P}[C_t = j \mid C_{t-1} = i], x_{1:t-1} \times \xi_{t-1}(i)\}$$

alors nous avons d'abord, en séparant le dernier terme du max, que

$$\begin{aligned}
\xi_{t+1}(j) &\stackrel{\text{def}}{=} \max_{c_{1:t}} \mathbb{P}[x_{1:t+1}, c_{1:t}, C_{t+1} = j], \\
&= \max_{i=1,\dots,K} \max_{c_{1:t-1}} \mathbb{P}\big[x_{1:t}, x_{t+1}, c_{1:t-1}, C_t = i, C_{t+1} = j \big], \\
&= \star.
\end{aligned}$$

Il ne reste maintenant qu'à isoler $\xi_t(i)$ et la probabilité qui ne dépend pas du fait que $C_t = i$, comme nous l'avons fait en $t = 1$.

$$\begin{aligned}
\star &= \max_{i=1,\dots,K} \max_{c_{1:t-1}} \{\mathbb{P}[x_{1:t},c_{1:t-1},C_t = i]\mathbb{P}[x_{t+1},C_{t+1} = j \mid x_{1:t},c_{1:t-1},C_t = i]\},\\
&= \max_{i=1,\dots,K} \max_{c_{1:t-1}} \left\{ \mathbb{P}[x_{1:t},c_{1:t-1},C_t = i]\mathbb{P}[C_{t+1} = j \mid C_t = i,x_{1:t}] \underbrace{\mathbb{P}[x_{t+1} \mid x_{1:t},C_{t+1} = j]}_{\text{Indépendant de } i} \right\}, \qquad \text{(III.6.2)}\\
&= \mathbb{P}[x_{t+1} \mid x_{1:t},C_{t+1} = j] \max_{i=1,\dots,K} \left\{ \mathbb{P}[x_{1:t},c_{1:t-1},C_t = i] \underbrace{\max_{c_{1:t-1}} \mathbb{P}[C_{t+1} = j \mid C_t = i,x_{1:t}]}_{\xi_t(i)} \right\},\\
&= \mathbb{P}[x_{t+1} \mid x_{1:t},C_{t+1} = j] \max_{i=1,\dots,K} \{\mathbb{P}[x_{1:t},c_{1:t-1},C_t = i]\xi_t(i)\}.
\end{aligned}$$

Puisque la relation est vraie en $t = 1$ et que si la relation est vraie en t, alors elle est vraie en $t+1$, il est possible d'affirmer par le principe d'induction que la relation tient $\forall t = 1,\dots,T$ et $\forall\, T \in \mathbb{N}^\star$. □

III.6.2. Démonstration de l'algorithme 5 : Algorithme de Viterbi (p.146)

Nous allons prouver par induction les relations II.2.18 et II.2.19 de l'algorithme 5.

Démonstration. Tout d'abord, nous avons en $t = T$ que

$$\begin{aligned}
i_T &\overset{\text{def}}{=} \arg\max_{c_T} \left\{ \max_{c_{1:T-1}} \mathbb{P}[x_{1:T},c_{1:T}] \right\},\\
&= \arg\max_{i=1,\dots,K} \left\{ \max_{c_{1:T-1}} \mathbb{P}[x_{1:T},c_{1:T-1},C_T = i] \right\},\\
&= \arg\max_{i=1,\dots,K} \xi_T(i).
\end{aligned}$$

En supposant maintenant que la relation tient jusqu'en $t+1$, ce signifie que nous connaissons les valeurs de $i_{t+1:T}$, nous avons que

$$
\begin{aligned}
i_t &\stackrel{\text{def}}{=} \arg\max_{c_t} \left\{ \max_{c_{1:t-1},c_{t+1:T}} \mathbb{P}[x_{1:T},c_{1:T}] \right\}, \\
&= \arg\max_{i=1,\ldots,K} \left\{ \max_{c_{1:t-1}} \mathbb{P}[x_{1:T},c_{1:t-1},C_t = i],C_{t+1:T} = i_{t+1:T} \right\}, \\
&= \arg\max_{i=1,\ldots,K} \left\{ \max_{c_{1:t-1}} \mathbb{P}[x_{1:t},c_{1:t-1},C_t = i] \mathbb{P}[x_{t+1:T},C_{t+1:T} = i_{t+1:T} \mid c_{1:t-1},C_t = i,x_{1:t}] \right\}, \\
&= \arg\max_{i=1,\ldots,K} \left\{ \max_{c_{1:t-1}} \mathbb{P}[x_{1:t},c_{1:t-1},C_t = i] \overbrace{\mathbb{P}[C_{t+1} = i_{t+1} \mid C_t = i,x_{1:t}]}^{\text{Indépendant de } c_{1:t-1}} \right. \\
&\qquad \left. \overbrace{\times \cancel{\mathbb{P}[x_{t+1:T},C_{t+2:T} = i_{t+2:T} \mid C_{t+1} = i_{t+1},x_{1:t}]}}^{\text{Ne dépendant pas de } c_{1:t}} \right\}, \\
&= \arg\max_{i=1,\ldots,K} \left\{ \mathbb{P}[C_{t+1} = i_{t+1} \mid C_t = i,x_{1:t}] \times \max_{c_{1:t-1}} \mathbb{P}[x_{1:t},c_{1:t-1},C_t = i] \right\}, \\
&= \arg\max_{i=1,\ldots,K} \{ \mathbb{P}[C_{t+1} = i_{t+1} \mid C_t = i,x_{1:t}] \xi_t(i) \}.
\end{aligned}
$$

Puisque la relation est vraie en $t = T$ et que si la relation est vraie en $t+1$, alors elle est vraie en t, il est possible d'affirmer par le principe d'induction que la relation tient $\forall t = T,\ldots,1$ et $\forall\, T \in \mathbb{N}^\star$. □

Bibliographie

Augustyniak, M. (jan. 2018). *STT6705V : Méthodes d'inférence pour les modèles à chaîne de Markov cachée [Notes de cours]*.

Baum, Leonard E. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes". In : *Inequalities III : Proceedings of the Third Symposium on Inequalities*. Sous la dir. d'Oved Shisha. University of California, Los Angeles : Academic Press, p. 1-8.

Casella, George et Roger Berger (juin 2001). *Statistical Inference*. Duxbury Resource Center. isbn : 0534243126.

Celeux, Gilles (1998). "Bayesian inference for mixture : The label switching problem". In : *Compstat*. Springer, p. 227-232.

Diebolt, Jean et Christian P. Robert (1994). "Estimation of Finite Mixture Distributions through Bayesian Sampling". In : *Journal of the Royal Statistical Society. Series B (Methodological)* 56.2, p. 363-375. issn : 00359246. url : http://www.jstor.org/stable/2345907.

Ferguson, T. S. (1996). "A Course in Large Sample Theory". In.

Forney, G. D. (1973). "The viterbi algorithm". In : *Proceedings of the IEEE* 61.3, p. 268-278.

Grimmett, G.R. et D.R. Stirzaker (2001). *Probability and random processes*. T. 80. 391. Oxford university press. url : http://scholar.google.com/scholar.bib?q=info:xzStZXK20NkJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&ct=citation&cd=0.

Hamilton, James D. (jan. 1994). *Time Series Analysis*. 1re éd. Princeton University Press. isbn : 0691042896. url : http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0691042896.

Hathaway, Richard J. (juin 1985). "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions". In : *Ann. Statist.* 13.2, p. 795-800. doi : 10.1214/aos/1176349557. url : https://doi.org/10.1214/aos/1176349557.

Jasra, Ajay, Chris C Holmes et David A Stephens (2005). “Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling”. In : *Statistical Science*, p. 50-67.

Kim, Chang-Jin (fév. 1994). “Dynamic Linear Model with Markov Switching”. In : *Journal of Econometrics* 60, p. 1-22. doi : 10.1016/0304-4076(94)90036-1.

Lessard, Sabin (2014). *Processus stochastiques : Cours et exercices corrigés*. Références sciences. Ellipses. isbn : 9782340002142. url : https://books.google.ca/books?id=FrC9rQEACAAJ.

Lystig, Theodore C et James P Hughes (2002). “Exact Computation of the Observed Information Matrix for Hidden Markov Models”. In : *Journal of Computational and Graphical Statistics* 11.3, p. 678-689. doi : 10.1198/106186002402. eprint : https://doi.org/10.1198/106186002402. url : https://doi.org/10.1198/106186002402.

Puolamäki, Kai et Samuel Kaski (2009). “Bayesian Solutions to the Label Switching Problem”. In : *Advances in Intelligent Data Analysis VIII*. Sous la dir. de Niall M. Adams, Céline Robardet, Arno Siebes et Jean-François Boulicaut. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 381-392. isbn : 978-3-642-03915-7.

Redner, Richard A. et Homer F. Walker (1984). “Mixture Densities, Maximum Likelihood and the Em Algorithm”. In : *SIAM Review* 26.2, p. 195-239. issn : 00361445. url : http://www.jstor.org/stable/2030064.

Seneta, E. (2006). *Non-negative Matrices and Markov Chains*. Springer Series in Statistics. Springer New York. isbn : 9780387327921. url : https://books.google.ca/books?id=J3bsjqQBCZUC.

Stephens, Matthew (2000). “Dealing with Label Switching in Mixture Models”. In : *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 62.4, p. 795-809. issn : 13697412, 14679868. url : http://www.jstor.org/stable/2680622.

Viterbi, A. (1967). “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In : *IEEE Transactions on Information Theory* 13.2, p. 260-269.

Zucchini, Walter et Iain Macdonald (avr. 2009). *Hidden Markov Models for Time Series : An Introduction Using R*. isbn : 9781420010893. doi : 10.1201/9781420010893.